Combining Information from Multiple Complex Surveys

By

Qi Dong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in The University of Michigan
2012

Doctoral Committee:

Associate Professor Michael R. Elliott, Co-chair
Professor Trivellore E. Raghunathan, Co-chair
Professor Richard Valliant
Assistant Professor Lu Wang
Professor Nathaniel Schenker, University of Maryland

DEDICATION

To My Parents and Grandparents

To Ye

To All of My Teachers

ACKNOWLEDGEMENTS

I would like to express my gratitude to all who have made this dissertation possible and because of whom my graduate experience has been one that I will cherish forever.

My deepest gratitude is to my advisors, Dr. Trivellore Raghunathan and Dr. Michael Elliott. I am grateful to them for their invaluable supervision of and guidance on this dissertation and every one of my steps towards completing this dissertation. I am also thankful to them for holding me to a high research standard and teaching me how to do research. Special thanks go to Dr. Michael Elliott for carefully reading and providing all the insightful comments on this dissertation.

I am indebted to my committee members, Dr. Nathaniel Schenker, Dr. Richard Valliant and Dr. Lu Wang for their insightful comments and constructive criticisms. I am grateful for the faculties in MPSM and JPSM for their support and instructions. James Lepkowski and Steve Heeringa have always been there when I need career instructions. Mick Couper was my academic advisor and could always be counted on to provide suggestions on my academic and career development. I am also thankful to my fellow students and staff in MPSM and JPSM.

This thesis is dedicated to my mother Rongrong Zhao, my father Jian Dong, my grandparents Yulan Cui and Fameng Dong, my aunts Xueping Dong, Xuehong Dong and Xuemei Dong and my wife Ye He. Their constant and unconditional love has always been the greatest support that makes all of this happen.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Increasingly, many substantive research questions require a degree of information not adequately collected in a single survey. Fortunately, survey organizations often repeatedly draw samples from the same population for different surveys and collect data on a considerable number of overlapping variables. This dissertation presents a new method for combining multiple surveys from a missing data perspective. Two major improvements of the new method include: 1) adjusting for the incompatibility among different sample designs and 2) combining an unlimited number of surveys.

The basic proposal is to simulate synthetic populations from which the respondents of each survey have been selected. In this process, different sampling designs of the multiple surveys will be taken into account. Once we have the synthetic populations, we could treat them as simple random samples with no complex sampling design features and borrow information across surveys to adjust for nonsampling errors or fill in the variables that are lacking in one or more surveys. Then, we can analyze each synthetic population with standard complete-data software for simple random samples and obtain valid inference by combining the point and variance estimates, first across synthetic populations within each survey, and then across multiple surveys. Existing methods borrowed from the disclosure risk field will be used to combine the synthetic populations from one survey; combining these results across multiple surveys will require the methods developed in this dissertation.

The first study develops the combining rule when multiple surveys present and proposes a model-based method to impute the unobserved population. The 2006 National Health Interview Survey (NHIS) and Medical Expenditure Panel Study (MEPS) are combined to estimate health insurance coverage. The second study develops a nonparametric method to impute the unobserved population, which is used to generate synthetic populations for the 2006 NHIS and MEPS and produce combined estimates of health insurance coverage. The third study extends the new method to combine surveys with missing variables. A new two-stage combining rule is developed to account for the uncertainty due to simultaneously imputing the missing variables and generating synthetic populations. The 2006 Behavioral Risk Factor Surveillance System (BRFSS) is combined with the NHIS and MEPS to estimate health insurance coverage.

CHAPTER 1

INTRODUCTION

1.1 Objectives

Increasingly many substantive research questions require a degree of information not adequately collected in a single survey. Fortunately, survey organizations often repeatedly draw samples from the same population for different surveys and collect data on a considerable number of overlapping variables. In the past decade, many statistical methods have been developed to combine information from multiple (mostly two) surveys allowing for improved inference. The existing combining survey methods have produced improved inference and achieved part of the following goals: 1) to reduce biases of the estimates from individual surveys due to sampling and/or nonsampling error (noncoverage error, nonresponse error and measurement error); 2) to increase precision of estimates from individual surveys by using the information from other surveys;

and 3) to produce a complete data set with all variables of interest by borrowing information across surveys and filling in the missing variables in individual surveys.

The objectives of this dissertation are: 1) to develop a new method for combining any number of surveys that adjusts for the incomparability among different data sources – the complex sampling design features; and 2) to combine the 2006 National Health Interview Survey (NHIS), Medical Expenditure Panel Study (MEPS) and Behavioral Risk Factor Surveillance System (BRFSS) and to estimate the US population's health insurance coverage.

The proposal is to simulate synthetic populations from which the respondents of each survey have been selected. In this process, different sampling designs of multiple surveys will be taken into account. Once we have the synthetic populations, we can treat them as simple random samples with no complex sampling design features and borrow information across surveys to adjust for nonsampling errors or fill in the variables that are lacking in one or more surveys. Then, we can analyze each synthetic population with standard complete-data software for simple random samples. And inference on the population quantity of interest can be obtained by combining the point and variance estimates first across synthetic populations within each survey using the existing combining rules for synthetic data, and then across multiple surveys using the methods developed in this dissertation.

1.2 Organization of this Dissertation

This dissertation is organized as follows: Chapter 2 presents the new combining survey method and develops the combining rule when multiple surveys are present. A

model-based method is proposed to impute the unobserved population and adjust for the complex sampling design features, which is then evaluated under two situations when the underlying model is linear and when the underlying model is log-linear. Finally, we apply the new combining survey method to combine the 2006 NHIS and MEPS and to estimate the US population's health insurance coverage.

To protect against model misspecification of the model-based method, Chapter 3 develops a nonparametric counterpart to impute the unobserved population and adjust for the complex sampling design features. We use the well-developed Bayesian bootstrap to adjust for stratification and clustering as well as the finite population Bayesian bootstrap (FPBB) to adjust for the unequal probability of selection. We provide both a theoretical proof and a simulation study to verify the point estimates from synthetic populations generated by the nonparametric method are unbiased and the variance estimates simulate the actual sampling variance. Finally, we apply the nonparametric method to generate synthetic populations for the 2006 NHIS and MEPS and use the new combining survey method in Chapter 2 to estimate health insurance coverage.

Chapter 4 extends the new combining survey method to the situation where there are missing variables in one or more surveys and we have to combine multiple surveys to obtain a complete list of variables of interest. A new two-stage combining rule is developed to account for the uncertainty due to simultaneously generate synthetic populations and impute the missing variables. We also conduct a simulation study to evaluate the two-stage combining rule. Finally, we apply the generalized method to first fill in the missing information in the 2006 BRFSS and then combine it with the NHIS and

MEPS to estimate health insurance coverage. Chapter 5 concludes the dissertation with discussions and describes the directions for future research.

CHAPTER 2

COMBINING INFORMATION FROM MULTIPLE COMPLEX SURVEYS

This chapter describes the use of multiple imputation to combine information from multiple surveys of the same underlying population. The basic proposal is to simulate synthetic populations from which the respondents of each survey have been selected. In this process, different sampling designs of the multiple surveys will be taken into account. We can then analyze each synthetic population with standard complete-data software for simple random samples and obtain valid inference by combining the point and variance estimates, first across synthetic populations within each survey using the existing combining rules for synthetic data, and then across multiple surveys using the methods developed in this chapter. A model-based method to produce the synthetic populations is discussed and evaluated. It is shown that the method in this chapter combines information from multiple surveys and produces more accurate and precise estimates for the statistics of interest.

2.1 Introduction

Survey agencies often repeatedly draw samples from the same or similar populations for different surveys and collect similar variables, sometimes even using the same frame. For example, the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES) are both conducted by the U.S. National Center for Health Statistics. These two surveys have similar target populations - the U.S. non-institutionalized population - and have a considerable overlap of questions. By combining information from multiple surveys, we hope to obtain more accurate inference for the population and be able to perform a variety of more comprehensive analysis than if we use the data from a single survey.

One of the biggest challenges in combining survey area is the comparability among multiple data sources. Surveys could use different sampling designs or modes of data collection, which may result in various sampling and nonsampling error properties, or surveys could ask the same question in different contexts or even for different reference periods. Instead of directly pooling the data from multiple surveys for a simple analysis, we need to adjust for the discrepancies among the data to make them comparable.

For example, suppose that two surveys have the same underlying population and the goal is to estimate the population mean, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$. Suppose one survey uses equal probability sampling (epsem) and the other one probability proportional to size sampling (PPS), and further that, for the second survey, both the variable of interest and the probability of selection are proportional to the measure of size, i.e., $Y_i, \pi_i \propto M_i$. The

estimate of the mean obtained under PPS sampling will have a much lower mean square error than equal probability sampling (Hansen and Hurwitz 1943; Jebe 1952). From an efficiency standpoint, the estimate from the PPS sample should be weighted more than the estimate from the SRS sample when combining those two surveys. Another example is combining data obtained from a face-to-face survey and a telephone survey, in which the undercoverage error of the telephone survey must be adjusted to account for the sampling frame excluding households without landline telephones (Raghunathan *et al.* 2007).

Various methods for combining data collected in two surveys have been proposed in the survey methodology literature (Hartley 1974; Skinner and Rao 1996; Elliott and Davis 2005; Raghunathan *et al*. 2007; Schenker *et al*. 2002, 2007, 2009). The most recent papers by Raghunathan *et al*. (2007) and Schenker *et al*. (2009) applied model-based approaches. The basic idea for the model-based approaches is to fit an imputation model to the data of better quality and use the fitted model to impute the values in the other samples of lower quality. As long as the imputation model is correctly specified, this approach can take advantage of the strengths of the multiple data sources and improve the statistical inference. However, as suggested by Reiter *et al.* (2006), when the sample is collected using complex sampling designs, ignoring those features could result in biased estimates from the design-based perspective. However, fully accounting for the complex sampling design features in practice is very difficult. For example, both Raghunathan *et al*. (2007) and Schenker *et al*. (2009) used a simplified method to adjust for stratification and clustering. Raghunathan *et al*. used a rudimentary concept of design effect and Schenker *et al*. used propensity scores to create adjustment subgroups for modeling. Both

of the papers could be improved if the complex sampling design features are better accounted for.

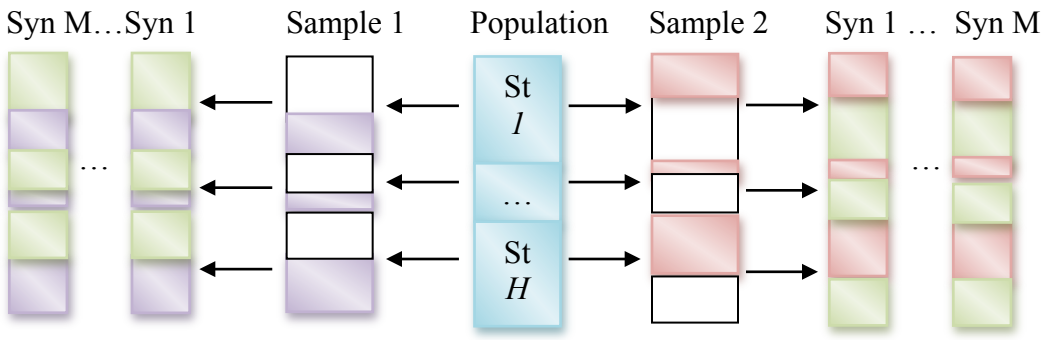
This chapter proposes a new method for combining multiple surveys from a missing data point of view that adjusts for the complex sampling design features in each survey. The unobserved population in each survey will be treated as missing data to be multiply imputed. The imputation model will account for complex design features. For each survey, the observed data and the multiply imputed unobserved population produce multiple synthetic populations. Once the whole population is generated, the complex sampling design features such as stratification, clustering and weighting will be of no use in the analysis and the synthetic populations can be treated as equivalent simple random samples. Finally, the estimate for the population quantity of interest will be calculated from each synthetic population and then will be combined first within each individual survey and then across multiple surveys.

This chapter proceeds as follows: Section 2.2 provides an overview of the proposed method. Section 2.3 discusses generating synthetic population while accounting for complex sampling design features using a model based method. Section 2.4 describes methodology to produce combined estimates from these multiple synthetic populations. Section 2.5 provides the results of a simulation study that shows the proposed method provides a more precise estimate of population mean than the estimate from any single survey. In Section 2.6, we apply the proposed method to combine the 2006 NHIS and the Medical Expenditure Panel Survey (MEPS) to estimate the health insurance coverage rates of the US population. Finally, Section 2.7 concludes with discussion and directions for future research.

2.2 Overview of Method

Figure 2.1 is an overview of the proposed method, in which we have two surveys covering the same underlying population with H strata. The samples (denoted by color cells) are drawn using different sampling designs. This chapter proposes to fill in the unobserved population (denoted by the blank cells) by building an imputation model based on the observed data of each survey. The multiple synthetic populations for Survey 1 and Survey 2 could be analyzed as simple random samples. We then will estimate the population quantity of interest from each synthetic population and combine them first within surveys then across surveys to produce the combined estimate. In the imputation model, we could use additional design variables that are available for the entire population. These variables are excluded here and from the figures and related formulas in the following sections for simplicity of exposition.

Figure 2.1 Illustration of data obtained using different sampling designs



2.3 Generating Synthetic Populations from Single Survey Data that Accounts for Complex Sampling Designs

We propose to generate synthetic populations using Bayesian finite population inference. The basic concept of Bayesian finite population inference involves imputing the non-sampled values of the population from the posterior predictive distribution based on the observed data. Assume the population values are $Y = (Y_1, \dots, Y_N)$ and the observed data, $Y_{obs} = (y_1, \dots, y_n)$ is obtained in a survey with sampling indicators $I = (I_1, \dots, I_N)$. Denote the population quantity of interest as $Q(Y)$. The Bayesian population inference allows for the use of parametric model $Pr(Y|\theta)$ for population data based on the posterior predictive distribution for the unobserved elements of the population $Pr(Y_{nob}|Y_{obs})$:

$$Pr(Y_{nob}\,|Y_{obs}) = \frac{Pr(Y)}{Pr(Y_{obs})}$$

$$= \frac{\int Pr(Y\,|\theta)\, Pr(\theta)\, d\theta}{Pr(Y_{obs})}$$

$$= \frac{\int Pr(Y_{nob}|Y_{obs}, \theta)\, Pr(Y_{obs}\,|\theta)\, Pr(\theta)\, d\theta}{Pr(Y_{obs})}$$

$$= \int Pr(Y_{nob}|Y_{obs}, \theta)\, Pr(\theta|Y_{obs})\, d\theta$$

(Ericson 1969; Holt and Smith 1979; Little 1993; Rubin 1987; Scott 1977; Skinner, Holt and Smith 1989). Here we use the model $Pr(Y|\theta)$ to approximate the entire population distribution $\Pr(Y)$ and average over the posterior distribution based on the sampled data $\Pr(\theta|Y_{obs})$. In the case that there are design variables known for the entire population available, the above model can be naturally extended by conditioning on these variables.

In the derivation of the posterior predictive distribution, we ignore the sampling indicator I. This requires:

1. $Pr(I|Y) = Pr(I|Y_{obs})$
2. $Pr(Y_{nob}|Y_{obs}, I, \theta) = Pr(Y_{nob}|Y_{obs}, \theta)$

Condition 1 requires the sampling design is ignorable (Rubin 1987), which is usually satisfied in probability samples. Condition 2 requires a model for the data $Pr(Y|\theta)$ that takes into consideration the complex sampling design features and is robust enough to capture all aspects of the distribution of Y.

Next we will illustrate two applications when the underlying model is linear and when the underlying model is log-linear. Other situations can be dealt with using a similar approach.

2.3.1 Linear Model

Suppose the population quantity of interest $Q(Y)$ is related to a normally distributed variable Y (possibly transformed from the original scale for normality). A simple random sample is drawn from the population, for which we measure the values of Y denoted by Y_{obs}. Further suppose there is a set of design variables $Z = (Z_{obs}, Z_{nob})$ known for the entire population on which we can regress Y using a linear model. For the observed data, we have $Y_{obs} = Z_{obs}\beta + e$, where β is a set of coefficients that relate the mean of Y to the covariance matrix Z_{obs}, e is the error term and has a multivariate normal distribution with mean zero and variance $\sigma^2 I$ and I is an identity matrix. Suppose $\theta = (\beta, log\sigma)$ has a uniform prior distribution over the appropriate dimensional real space. We fit the model using the sample data. Let $B = ({Z_{obs}}^t Z_{obs})^{-1} {Z_{obs}}^t Y_{obs}$ be the estimated regression coefficients, $SSE = (Y_{obs} - Z_{obs}B)^t (Y_{obs} - Z_{obs}B)$ be the residual sum of squares and df be the residual degrees of freedom. Assume T is the Cholesky

decomposition such that $TT^t = ({Z_{obs}}^t Z_{obs})^{-1}$. The relevant posterior distributions can be derived easily (Gelman, Carlin, Stern and Rubin, 2004), and the following steps draw the unobserved values from the posterior predictive distribution.

1. Generate a chi-square random deviate u with df degrees of freedom and define $\sigma_*^2 = SSE/u$.
2. Generate a vector $x = (x_1, \ldots, x_p)$ of dimension $p = rows(B)$ of random normal deviates and define $\beta_* = B + \sigma_* Tx$.
3. Let U denote the covariate matrix of the missing Y values. The unobserved values are $Y_{nob} = Z_{nob}\beta_* + \sigma_* v$, where v is an independent vector of random normal deviates.

When stratification, clustering and weighting are present, the model cannot be specified in such a simple way. For example, to account for clustering, we could include contextual variables such as county-level indicator in the imputation models. Or we could include random effects for the clusters and specify the covariance matrix correctly to capture the intracluster correlations. However, sometimes, we do not have sufficient information or large enough sample to fit such models. We propose a simplified and approximate approach to impute the unobserved values while adjusting for the complex sampling design features.

1. Estimate coefficients and covariance matrix:

Let B be the maximum likelihood estimates of β and V its asymptotic covariance matrix after complex design features are taken into account. V could be calculated using Tayor Linearization method or replication method to account for stratification, clustering and unequal probabilities of selection.

2. Approximate the posterior distribution of the coefficients:

Let T be the Cholesky decomposition such that $TT^t = V$. Generate a vector $x = (x_1, \ldots, x_p)$ of dimension $p = rows(B)$ of random normal deviates and define $\beta_* = B + Tx$.

3. Impute the unobserved values of the population as $Y_{nob} = Z_{nob}\beta_*$.

This approach results only in approximate draws from the posterior predictive distribution (Little and Rubin, 2002) as the draws for the parameter β are from the asymptotic approximation of its actual posterior distribution (Huber, 1967). Since the complex sampling design features are taken into account when we estimate the point and covariance estimates for the coefficients, the approximate posterior distribution of β reflects the distribution of the coefficients. The synthetic populations generated from the $i.i.d.$ draws simulate the underlying population.

2.3.2 Log-linear Model

A situation that appears frequently in survey data is the analyzing of a multidimensional contingency table since most of the variables collected in surveys are categorical. For simplicity of exposition, we assume Z and Y are both categorical, which create a two dimensional table. Assume Y is the variable of our interest with m levels; Z is a design variable with n levels (e.g., gender, race, etc) whose marginal distribution is known for the population. Assume $\pi_{ij}, i = 1, \dots, m, j = 1, \dots n$, represents the cell proportion of the ij^{th} cell, $\sum_{i=1}^{m}\sum_{j=1}^{n} \pi_{ij} = 1$. For this $m * n$ contingency table, the goal is to model the joint distribution between Z and Y for the actual data and use this model to generate the synthetic populations. The log-linear model has been developed to analyze multi-dimensional contingency tables (Agresti, 2002). Here, we have the following fully saturated model:

$$log(\pi_{ij}) = \lambda_0 + \lambda_i^Z + \lambda_j^Y + \lambda_{ij}^{ZY}, i = 1, \dots, m, j = 1, \dots n,$$

where $log(\pi_{ij})$ is the log of the probability that one observation falls in cell ij of the

contingency table, λ_i^Z is the main effect for Z, λ_j^Y is the main effect for Y and λ_{ij}^{ZY} is the interaction effect for Z and Y. We assume λ_i^Z, λ_j^Y and λ_{ij}^{ZY} are column vectors.

The above model includes all possible one-way and two-way effects and thus is saturated as it has the same number of effects as cells in the contingency table. The expected cell frequencies will always exactly match the observed frequencies, with no degrees of freedom remaining (Knoke and Burke, 1980). To avoid over fitting the data, we can consider lower dimensional models that exclude some or all of the interaction terms. We choose the model based on likelihood ratio tests or AIC or BIC criteria.

Following the idea for the linear model situation, the synthetic populations can be generated from the posterior predictive distribution from the model. However, when the data is collected under a complex sampling design, there is no existing statistical software that can produce both the point estimate and covariance estimate of the regression coefficients. We have to use replication method to adjust for stratification, clustering and weighting. Specifically, the synthetic populations can be generated from the following steps:

1. Estimate coefficients and covariance matrix:

Under the selected model (assume the saturated model here just for illustration), estimate the coefficients $\lambda = (\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY})', i = 1, \dots, m-1, j = 1, \dots n-1$ and the covariance matrix of the estimates $\hat{\lambda} = \left(\widehat{\lambda_0}, \widehat{\lambda_i^Z}, \widehat{\lambda_j^Y}, \widehat{\lambda_{ij}^{ZY}}\right)'$ after taking into account the complex design features using jackknife repeated replication (JRR).

- For each replication, withdraw one cluster, and inflate the weights for the respondents in the other clusters within the same stratum by $c_h/(c_h - 1)$ (replication weights), where c_h denotes the number of clusters within stratum h. Assume we have $\sum_{h=1}^{H} c_h = C$ clusters in total, then we have C replications. For each replication, we fit the log-linear model and obtain the maximum likelihood estimates (MLE) of the coefficients, $\lambda = (\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY})', i = 1, \dots, m-1, j = 1, \dots n-1$.

- For each replication, we use the replication weights to fit the log-linear model. Specifically, we use the replication weights to calculate the size of each cell of the contingency table, which is used to fit the log-linear model. We denote the MLE for the r^{th} replication by a column vector, $\widehat{\lambda_r}, r = 1, \dots, c_h$ for stratum h. Notice that $\lambda = (\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY})', i = 1, \dots, m-1, j = 1, \dots n-1$ is a mn by 1 column vector. We denote $\lambda = (\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY})' = (\lambda_0, \lambda_1, \dots, \lambda_{\mathrm{mn}})'$. Similarly, $\widehat{\lambda_r}, r = 1, \dots, c_h, h = 1, \dots, H$ are also mn by 1 column vectors denoted by $(\widehat{\lambda_0}^{<r>}, \widehat{\lambda_1}^{<r>}, \dots, \widehat{\lambda_{\mathrm{mn}}}^{<r>})'$.

The MLE of the coefficients $\lambda = (\lambda_0, \lambda_i^Z, \lambda_j^Y, \lambda_{ij}^{ZY})', i = 1, \dots, m-1, j = 1, \dots n-1$ can be obtained by $\widehat{\lambda_{MLE}} = \sum_{h=1}^{H} \sum_{r=1}^{c_h} \widehat{\lambda_r} / C$. For the mn by mn covariance matrix, the jackknife replication estimate of the pq^{th} $(p, q = 1, \dots, mn)$ element is the covariance between the $p^{th}$and q^{th} coefficients, which is given by:

$$\sum_{h=1}^{H} \frac{c_h - 1}{c_h} \sum_{\mathrm{r}=1}^{c_h} \left(\widehat{\lambda_{\mathrm{p}}}^{<r>} - \overline{\hat{\lambda}_{\mathrm{p}}}\right)\left(\widehat{\lambda_{\mathrm{q}}}^{<r>} - \overline{\hat{\lambda}_{\mathrm{q}}}\right)$$

,where $\overline{\hat{\lambda}_{\mathrm{p}}} = \sum_{\mathrm{h}=1}^{\mathrm{H}} \sum_{\mathrm{r}=1}^{c_h} \widehat{\lambda_{\mathrm{p}}}^{<r>} / \mathrm{C}$ and $\overline{\hat{\lambda}_{\mathrm{q}}} = \sum_{\mathrm{h}=1}^{\mathrm{H}} \sum_{\mathrm{r}=1}^{c_h} \widehat{\lambda_{\mathrm{q}}}^{<r>} / \mathrm{C}$. This gives us the correct variance estimate of $\widehat{\lambda_{MLE}}$.

2. Approximate the posterior distribution of the coefficients:

Let T denote the Cholesky decomposition such that $TT^t = cov(\widehat{\lambda_{MLE}})$. Generate a vector z of random normal deviates and define $\Lambda_* = \widehat{\lambda_{MLE}} + Tz$.

3. Impute the unobserved values of the population:

Suppose L draws, $\Lambda_1, \dots, \Lambda_L$, are made from the approximate posterior distribution of λ. For each $l = 1, \dots, L, \Lambda_l = ({\Lambda_0}^{(l)}, \Lambda_i^{X(l)}, \Lambda_j^{Y(l)}, \Lambda_{ij}^{XY(l)})', i = 1, \dots, m-1, j = 1, \dots n-1$, we can generate one synthetic table using the assumed model: $log\left({\pi_{ij}}^{(l)}\right) = {\Lambda_0}^{(l)} + \Lambda_i^{X(l)} + \Lambda_j^{Y(l)} + \Lambda_{ij}^{XY(l)}, i = 1, \dots, m-1, j = 1, \dots n-1$. Once the cell proportions are determined, we can generate the synthetic table of any size.

4. Post-stratify/ Constraints on Margins:

Survey agencies usually post-stratify the collected data according to some auxiliary variables whose population margins are known. The post-stratification adjusts for the nonresponse and noncoverage error. If the imputation model does not approximate the population well, we could lose a fairly large amount of information, which may bias the estimates or inflate the variance estimates from the synthetic populations. As suggested by Raghunathan *et al.* (2003), we can constrain the marginal distribution of the design variables in the synthetic populations to match their marginal distributions in the population using the iterative proportional fitting (IPF) algorithm.

For example, suppose we denote the cell counts for the actual data, μ_{ij} and the l^{th} unconstrained synthetic data as $\hat{\mu}_{ij}^{l}, i = 1, \dots, n, j = 1, \dots, m, l = 1, \dots, L$ and further denote the margins of the actual data and synthetic populations as $\mu_{i+} = \sum_{j=1}^{m} \mu_{ij}$, $\hat{\mu}_{i+}^{l} = \sum_{j=1}^{m} \hat{\mu}_{ij}^{l}$, etc, where μ_{i+} and μ_{+j} are known for the population . We constrain the margins of the unconstrained synthetic populations using the following algorithm:

1. $\hat{\mu}_{ij}^{l\ (t)} = \hat{\mu}_{ij}^{l\ (t-1)} \frac{\mu_{i+}}{\hat{\mu}_{i+}^{l\ (t-1)}}$
2. $\hat{\mu}_{ij}^{l\ (t+1)} = \hat{\mu}_{ij}^{l\ (t)} \frac{\mu_{+j}}{\hat{\mu}_{+j}^{l\ (t)}}$

Step a and b are repeated until the fitted table converges, i.e., $\left|1 - \hat{\mu}_{i+}^{l\ (t)} / \hat{\mu}_{i+}^{l\ (t-1)}\right| + \left|1 - \hat{\mu}_{+j}^{l\ (t)} / \hat{\mu}_{+j}^{l\ (t-1)}\right| < c$, where c is the pre-determined criteria, usually a small number like 0.0001.

2.4 Combing Rule for the Synthetic Populations from Multiple Surveys

Assume that $Q = Q(Y)$ is the population quantity of interest that may depend upon the a set of variables Y which is collected in multiple surveys. It could be a population mean, proportion or total, a vector of regression coefficients, etc. For simplicity of exposition, in this chapter, Q is assumed to be a scalar and Y is assumed to be one variable. Suppose under some sampling design, the analyst would use a point estimate q and an associated measure of uncertainty v. For example, q could be the maximum likelihood estimate of Q and v could be the inverse of the Fisher information. Alternatively, the Bayesian approach would estimate q and v using the posterior mean and variance of Q based on the actual sample data observed. A frequentist could construct an unbiased estimate q of Q with v as its sampling variance.

Assuming that we create L synthetic populations, $\mathcal{P}_l$, $l = 1, \dots, L$, denote Q_l as the corresponding estimate of the population quantity Q obtained from synthetic population l, with U_l denoting the within-imputation variance of Q_l. For large samples or a large

number of synthetic populations when the sample size is small, the posterior variance of Q is

$$T_L = \left(1 + \frac{1}{L}\right)\frac{1}{L-1}\sum_{l=1}^{L}(Q_l - \bar{Q}_L)(Q_l - \bar{Q}_L)' - \frac{1}{L}\sum_{l=1}^{L} U_l$$

$$= \left(1 + \frac{1}{L}\right) B_L - \overline{U}_L \,, \qquad (1)$$

where B_L is the between-imputation variance, $\overline{U}_L$ is the average of the within-imputation variance and $\bar{Q}_L = \frac{1}{L}\sum_{l=1}^{L} Q_l$ is the mean of Q across the L synthetic populations (Raghunathan *et al.* 2003). Since Q_l is computed from the whole synthetic population, the within-imputation variance could be ignored in the calculation of T_L, i.e., expression (1) can be reduced to

$$T_L = (1 + 1/\mathrm{L})B_L \qquad (2)$$

(Raghunathan *et al.* 2003). From these results, the Monte Carlo method can be used to draw inferences for the population quantity of interest, Q. In practice, it is unrealistic to impute the whole population, which could be hundreds of millions of units. We only need to generate the size of the synthetic population large enough compared to the sample size so the within-imputation variance $\overline{U}_L$ can be ignored.

In the context of combining information from multiple surveys, I will need to generate L synthetic populations for each survey and combine the estimates within each survey, and then combine across all S surveys as well. Raghunathan *et al.* (2003) developed a combining rule for synthetic populations from a single survey. But this combining rule will not yield valid inference for the parameters of interest for multiple surveys, since the models to generate synthetic populations (the predictive distribution of

the unobserved values given the observed values, denoted by Raghunathan et al (2003) by $Pr(y_{nob}|y_{obs})$) for the multiple surveys are different. Thus, a new rule for combining estimates across multiple surveys needs to be developed.

Table 2.1 Glossary

Symbol	Notation
$\boldsymbol{Q}$	The population quantity of interest
$\boldsymbol{q_{obs}^{(s)}}$	The estimate of Q obtained from the observed data of survey s
$\boldsymbol{B_s}$	The actual sampling variance of survey s
$\boldsymbol{b_{obs}^{(s)}}$	The estimate of the actual sampling variance of survey s
$\boldsymbol{\mathcal{P}_l^{(s)}}$	The l^{th} synthetic population of survey s
$\boldsymbol{Q_l^{(s)}}$	Population quantity of interest based on $\mathcal{P}_l^{(s)}$
$\boldsymbol{\overline{Q_L^{(s)}} = \frac{1}{L}\sum_{l=1}^{L} Q_l^{(s)}}$	Population quantity of interest from the L synthetic populations of Survey s
$\boldsymbol{B_L^{(s)}} = \boldsymbol{\frac{1}{L-1}\sum_{l=1}^{L}\left(Q_l^{(s)} - \overline{Q_L^{(s)}}\right)\left(Q_l^{(s)} - \overline{Q_L^{(s)}}\right)'}$	The variance of the population quantity of interest from the L synthetic populations of Survey s
$\boldsymbol{\overline{q_L^{(s)}}}$	The estimate of $\overline{Q_L^{(s)}}$
$\boldsymbol{b_L^{(s)}}$	The estimate of $B_L^{(s)}$

Assume $\overline{Q_L^{(s)}}$ and $B_L^{(s)}$ are respectively the combined estimator of the population quantity of interest and its variance for Survey s obtained using the combining formulas for synthetic populations in a single survey setting (Raghunathan *et al.* 2003) (For notation definitions in this section, see Table 2.1.). The approach considers ($\overline{Q_L^{(s)}}$, $B_L^{(s)}$), $s = 1, \dots, S$, as sufficient summaries of the synthetic population $\mathcal{P}_{syn}^{(s)} = \{\mathcal{P}_l^{(s)}, l =$

$1, \ldots, L\}, s = 1, \ldots, S$. The goal is to approximate the posterior density of Q conditional on $\mathcal{P}_{syn}^{(s)}, s = 1, \ldots, S$ or equivalently, $(\overline{Q_L^{(s)}}, B_L^{(s)}), s = 1, \ldots, S$. To do this, we need to make three asymptotic distributional assumptions:

Assumption 1: The repeated sampling distribution of the observed data statistic for each survey, $Q_{obs}^{(s)}, s = 1, \ldots, S$, is normal with mean the population quantity Q and some sampling variance B_s, i.e., $Q_{obs}^{(s)} | Y \sim N(Q, B_s), s = 1, \ldots, S$.

Assumption 2: The posterior distribution of the population quantity of interest Q based on the synthetic populations generated from Survey s is approximately normally distributed with mean $q_{obs}^{(s)}$ and variance $b_{obs}^{(s)}$, where $b_{obs}^{(s)}$ is an estimate of the sampling variance of $q_{obs}^{(s)}$: $Q_l^{(s)} | \, y_{obs} \sim \; N(q_{obs}^{(s)}, b_{obs}^{(s)})$.

Assumption 3: For Survey s, the variance estimator obtained from the L synthetic populations $T_L^{(s)}$, is unbiased for B_s with negligible sampling variability, i.e., $B_s \approx T_L^{(s)} = \left(1 + \frac{1}{L}\right) B_L^{(s)} - \overline{U}_L^{(s)}$. Since the whole population is generated, $\overline{U}_L^{(s)} = 0$, which means $B_s \approx \left(1 + \frac{1}{L}\right) B_L^{(s)} \approx B_L^{(s)}$ (when L is large).

Assumption 1 can be satisfied for many statistics that follow Central Limit Theorem (e.g., means, pseudo maximum likelihood estimates) as long as surveys use probability sampling and the statistical inference takes the sampling design into account. *Assumption 2* can be satisfied by imputing the unobserved part of the population for each survey using a model that is consistent with respect to the design of that survey. In other words, the complex design features and different survey error properties need to be built into the imputation model $Pr\,(Y_{nob}|Y_{obs})$. *Assumption 3* is usually satisfied for large samples or for a large number of synthetic populations when the sample sizes are small.

2.4.1 Combining Rule when *L* is large

When L is large, we suggest approximating the posterior distribution of Q as a normal distribution with mean

$$\overline{q_L} = \frac{\sum_{s=1}^{S} \frac{\overline{q_L^{(s)}}}{b_L^{(s)}}}{\sum_{s=1}^{S} \frac{1}{b_L^{(s)}}}$$

and variance

$b_L = \frac{1}{\sum_{s=1}^{S} \frac{1}{b_L^{(s)}}}$. We derive this result as follows:

Suppose $\overline{Q_L^{(s)}}$ and $B_L^{(s)}$ are the combined estimator of the population quantity of interest and its variance estimator based on the $L(=\infty)$ synthetic populations for Survey s respectively. We denote $B_1, \dots, B_S$ as the sampling variance from the observed data of the S surveys (*Assumption 2*). We assume the sample size is reasonably large, the sampling distribution of the sample quantity of interest is approximately normally distributed and the approach to generate synthetic populations is consistent with the design of each survey. Thus the three assumptions above are satisfied. The goal here is to derive the posterior predictive distribution of the parameter of interest, Q given the synthetic populations from multiple surveys when the number of synthetic populations is large, *i.e.*, $\pi(Q, B_1, \dots, B_s | \mathcal{P}_l^{(s)}, l = 1, \dots, L, s = 1, \dots, S)$. Since the entire population is imputed, there is no within-imputation variance. Here we treat $(\overline{Q_L^{(s)}}, B_L^{(s)})$ as sufficient summaries of the synthetic population $\mathcal{P}_l^{(s)}, l = 1, \dots, L, s = 1, \dots, S$, so that the posterior predictive distribution can be written as $\pi(Q, B_1, \dots, B_s | \overline{Q_L^{(1)}}, B_L^{(1)}, \dots, \overline{Q_L^{(S)}}, B_L^{(S)})$.

From Bayes' Theorem,

$$\pi(\mathrm{Q}, B_1, \dots, B_S \mid \overline{Q_L^{(1)}}, B_L^{(1)}, \dots, \overline{Q_L^{(S)}}, B_L^{(S)}) \propto \pi(\overline{Q_L^{(1)}}, B_L^{(1)}, \dots, \overline{Q_L^{(S)}}, B_L^{(S)} \mid Q, B_1, \dots, B_S)$$

$$\pi(\mathrm{Q}, B_1, \dots, B_S),$$

where the first part is the likelihood and the second part is the prior distribution.

The derivation of the likelihood

From Raghunathan *et al.* (2003), we have the following approximate sampling distribution:

$$(\overline{Q_L^{(s)}} | B_L^{(s)}, \mathrm{Q}, B_s) \sim N(\mathrm{Q}, B_s)$$

$$(B_L^{(s)}/B_s | \mathrm{Q}, B_s) \sim \chi^2_{L-1}/(L-1), s = 1, \dots, S, \quad (3)$$

i.e., we can write the distributions as

$$\pi(\overline{Q_L^{(s)}} | B_L^{(s)}, Q, B_s) \propto B_s^{-1/2} \exp\left(-\frac{\left(\overline{Q_L^{(s)}} - \mathrm{Q}\right)^2}{2B_s}\right)$$

$$\pi(B_L^{(s)} \mid Q, B_s) \propto {B_L^{(s)}}^{(L-3)/2} B_s^{-(L-3)/2} exp\left(-\frac{B_L^{(s)}}{2B_s}\right), \text{ s=1, …, S.} \quad (4)$$

When the number of synthetic populations, L, is infinite, $\chi^2_{L-1}/(L-1)$ converges to 1, which implies $B_L^{(s)}/B_s \approx 1$ or, B_s can be approximated by $B_L^{(s)}$, i.e., $(\overline{Q_L^{(s)}} | B_L^{(s)}, Q, B_s) \sim N(\mathrm{Q}, B_L^{(s)})$.

Since each survey is conducted independently, we have

$$\pi(\overline{Q_L^{(1)}}, B_L^{(1)}, \dots, \overline{Q_L^{(S)}}, B_L^{(S)} \mid Q, B_1, \dots, B_S) = \prod_{s=1}^{S} \pi(\overline{Q_L^{(s)}}, B_L^{(s)} \mid \mathrm{Q}, B_s).$$

$$= \prod_{s=1}^{S} \pi(\overline{Q_L^{(s)}} | B_L^{(s)}, Q, B_s)\, \pi(B_L^{(s)} | Q, B_s)$$

$$= \prod_{s=1}^{S} \pi\left(\overline{Q_L^{(s)}} \middle| Q, B_L^{(s)}\right) \pi\left(B_L^{(s)} \middle| B_s\right)$$

$$\propto \exp\left(-\sum_{s=1}^{S} \frac{\left(\overline{Q_L^{(s)}} - \mathrm{Q}\right)^2}{2B_L^{(s)}}\right) \prod_{s=1}^{S} {B_L^{(s)}}^{(L-4)/2} {B_s}^{-(L-3)/2} exp\left(-\frac{B_L^{(s)}}{2B_s}\right),$$

where the third equation is because $\pi(\overline{Q_L^{(s)}} | B_L^{(s)}, Q, B_s)$ doesn't involve B_s once $B_L^{(s)}$ is known and $\pi(B_L^{(s)} | Q, B_s)$ doesn't involve Q.

The derivation of the posterior predictive distribution

We use a non-informative prior, $\pi(\mathrm{Q}, B_1, \ldots, B_S)$, i.e., $\pi(\mathrm{Q}, B_1, \ldots, B_S) \propto \prod_{s=1}^{S} {B_s}^{-1}$, though a weak conjugate prior leads to the same conclusion. The posterior predictive distribution is

$$\pi(\mathrm{Q}, B_1, \ldots, B_S \mid \overline{Q_L^{(1)}}, B_L^{(1)}, \ldots, \overline{Q_L^{(S)}}, B_L^{(S)})$$

$$\propto \pi(\overline{Q_L^{(1)}}, B_L^{(1)}, \ldots, \overline{Q_L^{(S)}}, B_L^{(S)} \mid Q, B_1, \ldots, B_S)\, \pi(\mathrm{Q}, B_1, \ldots, B_S)$$

$$\propto \exp\left(-\sum_{s=1}^{S} \frac{\left(\overline{Q_L^{(s)}} - \mathrm{Q}\right)^2}{2B_L^{(s)}}\right) \prod_{s=1}^{S} {B_L^{(s)}}^{(L-4)/2} {B_s}^{-(L-3)/2} exp\left(-\frac{B_L^{(s)}}{2B_s}\right) {B_s}^{-1} \qquad (5)$$

Then, the marginal posterior distribution of Q can be obtained as:

$$\pi(Q \mid \overline{Q_L^{(1)}}, B_L^{(1)}, \ldots, \overline{Q_L^{(S)}}, B_L^{(S)}) = \int \ldots \int \pi(\mathrm{Q}, \mathrm{B} \mid \overline{Q_L^{(1)}}, B_L^{(1)}, \ldots, \overline{Q_L^{(S)}}, B_L^{(S)})\ \mathrm{d}B_1 \ldots \mathrm{d}B_S$$

$$\propto \exp(-\sum_{s=1}^{S} \frac{\left(\overline{Q_L^{(s)}}-\mathrm{Q}\right)^2}{2B_L^{(s)}}) \int \dots \int \prod_{s=1}^{S} B_L^{(s)\frac{L-4}{2}} B_s^{-\frac{L-1}{2}} exp(-\frac{B_L^{(s)}}{2B_s}) \, \mathrm{d}B_1 \dots \mathrm{d}B_S \qquad (6)$$

Notice that the terms within the integration in the last line of expression (6) is the kernel of a Chi-square Distribution for $1/B_s$. From the equality

$$\int \dots \int \prod_{s=1}^{S} B_L^{(s)\frac{L-4}{2}} B_s^{-\frac{L-1}{2}} exp(-\frac{B_L^{(s)}}{2B_s}) \, \mathrm{d}B_1 \dots \mathrm{d}B_S = f(L, B_L^{(1)}, \dots, B_L^{(S)}),$$

we obtain,

$$\pi\left(\mathrm{Q}\middle|\overline{Q_L^{(1)}}, B_L^{(1)}, \dots, \overline{Q_L^{(S)}}, B_L^{(S)}\right)$$

$$\propto exp(-\sum_{s=1}^{S} \frac{\left(\overline{Q_L^{(s)}}-\mathrm{Q}\right)^2}{2B_L^{(s)}})$$

$$\propto exp(-(Q - \frac{\sum_{s=1}^{S} \frac{\overline{Q_L^{(s)}}}{B_L^{(s)}}}{\sum_{s=1}^{S} \frac{1}{B_L^{(s)}}})^2 / \frac{1}{\sum_{s=1}^{S} \frac{1}{2B_L^{(s)}}}). \qquad (7)$$

This implies that the posterior predictive distribution $\pi(Q \mid \overline{Q_L^{(1)}}, B_L^{(1)}, \dots, \overline{Q_L^{(S)}}, B_L^{(S)})$ can be approximated by a normal distribution with the following parameters:

$$N\left(\sum_{s=1}^{S} \frac{\overline{Q_L^{(s)}}}{B_L^{(s)}} / \sum_{s=1}^{S} \frac{1}{B_L^{(s)}}, 1/(\sum_{s=1}^{S} \frac{1}{B_L^{(s)}})\right). \qquad (8)$$

Assume an unbiased estimate for $Q_l^{(s)}$ from the synthetic population is $q_l^{(s)}$ and that $\overline{q_L^{(s)}}$ and $b_L^{(s)}$ are the estimates of $\overline{Q_L^{(s)}}$ and $B_L^{(s)}$. The combined estimate of Q will be $\sum_{s=1}^{S} \frac{\overline{q_L^{(s)}}}{b_L^{(s)}} / \sum_{s=1}^{S} \frac{1}{b_L^{(s)}}$ with variance estimate $1/(\sum_{s=1}^{S} \frac{1}{b_L^{(s)}})$.

2.4.2 T-corrected Distribution for Small/Moderate *L*

Thus far, we have implicitly assumed that the actual posterior distribution of Y_{nobs} for each survey could be simulated perfectly in the sense that inferences have been based on a large number of synthetic populations (large L). In practice, it is sometimes unrealistic to generate a large number of synthetic populations, especially when the sample size is so large that it is computationally intensive to impute the unobserved population. In this section, we modify the theory for small or moderate L (e.g., $L < 50$).

Below we show that, for finite L, the posterior distribution of Q follows a t distribution with mean $\overline{q_L}$, scale $(1 + L^{-1})b_L$ and degrees of freedom $(L-1)/\sum_{s=1}^{S}(\frac{\frac{1}{b_L^{(s)}}}{\sum_{s=1}^{S}\frac{1}{b_L^{(s)}}})^2$. Assume L synthetic populations are generated for survey $s, s = 1, \dots, S$, where L is small or moderate. Let $Q_l^{(s)}$ represent the estimator from the l^{th} synthetic population of Survey s, $l = 1, \dots, L$, *s=1, … , S*. Let $\overline{Q_L^{(s)}}$ and $B_L^{(s)}$ represent the combined estimator of the population quantity of interest and its variance estimator for survey $s, s = 1, \dots, S$. Let $\overline{Q_\infty}$, B_∞ represent the combined estimator across the S surveys when we have large or infinite number of synthetic populations, i.e., $\overline{Q_\infty} = \sum_{s=1}^{S} \frac{\overline{Q_\infty^{(s)}}}{B_\infty^{(s)}} / \sum_{s=1}^{S} \frac{1}{B_\infty^{(s)}}$ and $B_\infty = 1/(\sum_{s=1}^{S} \frac{1}{B_\infty^{(s)}})$. Let $\overline{Q_L}, B_L$ represent the combined estimator across the

S surveys when we have small or moderate number of synthetic populations, i.e., $\overline{Q_L} = \sum_{s=1}^{S} \frac{\overline{Q_L^{(s)}}}{B_L^{(s)}} / \sum_{s=1}^{S} \frac{1}{B_L^{(s)}}, B_L = 1/(\sum_{s=1}^{S} \frac{1}{B_L^{(s)}})$. The goal is to approximate the conditional distribution: $Pr(Q|\overline{Q_L^{(1)}}, B_L^{(1)}, \dots, \overline{Q_L^{(S)}}, B_L^{(S)})$ from the results for large or infinite L.

From *4.1*, the posterior distribution of Q is approximated as a normal distribution with mean $\overline{Q_\infty}$ and variance B_∞, i.e.,

$Q|\,\overline{Q_\infty}\,, B_\infty \sim N(\overline{Q_\infty}\,, B_\infty)$.

This can be also be written as:

$Q|\,\overline{Q_\infty}\,, B_\infty, \overline{Q_L}\,, B_L \sim N(\overline{Q_\infty}\,, B_\infty)$.

The sampling distribution of $\overline{\mathbf{Q_L}}$ Given $(\overline{\mathbf{Q_\infty}}\,, \mathbf{B_\infty})$

Within individual surveys, we have the following t-corrected distribution when the number of synthetic populations is small or moderate (Raghunathan et al. 2003),

$Q|\overline{Q_L^{(s)}}, B_L^{(s)} \sim t_{L-1}(\overline{Q_L^{(s)}}, (1 + 1/L)B_L^{(s)})$.

This implies:

$(Q|\overline{Q_L^{(s)}}, B_L^{(s)}, B_s) \sim N(\overline{Q_L^{(s)}},\ (1 + 1/L)B_s)$

$(B_L^{(s)}/B_s|\,\overline{Q_L^{(s)}}, B_s) \sim\ \ \chi^2_{L-1}/(L-1)$.

This implies when L goes to infinity, $B_\infty^{(s)} \approx\ B_s$. When L is small or moderate, we have $(Q|\overline{Q_L^{(s)}}, B_L^{(s)}, B_s) \sim N(\overline{Q_L^{(s)}},\ (1 + 1/L)B_s)$. This implies $B_L^{(s)} \approx\ (1 + 1/L)B_s) \approx (1 + 1/L)\ B_\infty^{(s)}$.

Within individual surveys, it is reasonable to suppose (Rubin 1987)

$Q_l^{(s)} | \overline{Q_\infty^{(s)}}, B_\infty^{(s)} \sim N(\overline{Q_\infty^{(s)}}, B_\infty^{(s)}), s = 1, \ldots, S$ that $\overline{Q_L^{(s)}}$ is the mean of L *i.i.d.* draws from this distribution.

The conditional distribution of $\overline{\mathbf{Q}_\infty}$ given $(\overline{\mathbf{Q}_\mathbf{L}}, \mathbf{B}_\mathbf{L}, \mathbf{B}_\infty)$

Since the normal sampling distribution of $Q_l^{(s)}, s = 1, \ldots, S$, we have $\overline{Q_L^{(s)}} = \frac{1}{L}\sum_{l=1}^{L} Q_l^{(s)} \mid \overline{Q_\infty^{(s)}}, B_\infty^{(s)} \sim N(\overline{Q_\infty^{(s)}}, B_\infty^{(s)}/L)$. If the prior distribution of $\overline{Q_\infty^{(s)}}$ conditional on $B_\infty^{(s)}$ is proportional to a constant, the conditional distribution of $\overline{Q_\infty^{(s)}}$ given $\overline{Q_L^{(s)}}$, $B_L^{(s)}$ and $B_\infty^{(s)}$ is normal:

$$\overline{Q_\infty^{(s)}} \mid \overline{Q_L^{(s)}}, B_L^{(s)}, B_\infty^{(s)} \sim N(\overline{Q_L^{(s)}}, B_\infty^{(s)}/L)$$

Since $B_L^{(s)} \approx \left(1 + \frac{1}{L}\right) B_\infty^{(s)}$, this leads to:

$$\overline{Q_\infty} = \sum_{s=1}^{S} \frac{\overline{Q_\infty^{(s)}}}{B_\infty^{(s)}} / \sum_{s=1}^{S} \frac{1}{B_\infty^{(s)}} \mid \overline{Q_L}, B_L, B_\infty \sim \mathrm{N}(\overline{Q_L}, B_\infty/\mathrm{L}),$$

where $\overline{Q_L} = \sum_{s=1}^{S} \frac{\overline{Q_L^{(s)}}}{B_L^{(s)}} / \sum_{s=1}^{S} \frac{1}{B_L^{(s)}}$ and $B_\infty = 1/\sum_{s=1}^{S} \frac{1}{B_\infty^{(s)}}$.

The conditional distribution of Q given $(\overline{\mathbf{Q}_\mathbf{L}}, \mathbf{B}_\mathbf{L}, \mathbf{B}_\infty)$

From $Q \mid \overline{Q_\infty}, B_\infty, \overline{Q_L}, B_L \sim N(\overline{Q_\infty}, B_\infty)$ and $\overline{Q_\infty} \mid B_\infty, \overline{Q_L}, B_L \sim \mathrm{N}(\overline{Q_L}, B_\infty/\mathrm{L})$, we have $Q \mid B_\infty, \overline{Q_L}, B_L$ follows a normal distribution with mean

$$E(Q \mid B_\infty, \overline{Q_L}, B_L) = E(E(Q \mid B_\infty, \overline{Q_L}, B_L \mid \overline{Q_\infty})) = E(\overline{Q_\infty} \mid B_\infty, \overline{Q_L}, B_L) = \overline{Q_L}$$

and variance

$V(Q|\, B_\infty, \overline{Q_L}, B_L) = V(E(Q|\, B_\infty, \overline{Q_L}, B_L|\overline{Q_\infty})) + E(V(Q|\, B_\infty, \overline{Q_L}, B_L|\overline{Q_\infty})) =$

$V(\overline{Q_\infty}|\, B_\infty, \overline{Q_L}, B_L) + E(B_\infty|B_\infty, \overline{Q_L}, B_L) = B_\infty/\mathrm{L} + B_\infty,$

i.e.,

$Q|\, B_\infty, \overline{Q_L}, B_L \sim N(\overline{Q_L}, (1 + L^{-1})B_\infty).$

The conditional distribution of B_∞ given B_L

$\frac{B_L}{B_\infty}|\, B_L = B_L/\frac{1}{\sum_{s=1}^{S}\frac{1}{B_\infty^{(s)}}} = B_L/\frac{1}{\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}\frac{B_L^{(s)}}{B_\infty^{(s)}}} \sim B_L/\frac{1}{\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}\frac{\chi^2_{L-1}}{L-1}} = B_L\left(\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}\frac{\chi^2_{L-1}}{L-1}\right) =$

$\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}\frac{\chi^2_{L-1}}{L-1} / \sum_{s=1}^{S}\frac{1}{B_L^{(s)}}.$

The term $(\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}\chi^2_{L-1})/\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}$ is a weighted sum of S chi-square distributions of the same degree of freedom, which can be approximated by $a * \chi_b^2$, where a and b are the parameters to be determined. If we denote $w_s = \frac{1}{B_L^{(s)}} / \sum_{s=1}^{S}\frac{1}{B_L^{(s)}}$, then $(\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}\chi^2_{L-1})/\sum_{s=1}^{S}\frac{1}{B_L^{(s)}} = \sum_{s=1}^{S} w_s\, \chi^2_{L-1}$.

By equating the first and second moments of $\sum_{s=1}^{S} w_s\, \chi^2_{L-1}$ and $a * \chi_b^2$, we obtain $a = \sum_{s=1}^{S} {w_s}^2 \; and \; b = (L-1)/\sum_{s=1}^{S} {w_s}^2$.

Thus, $B_L/B_\infty|\, B_L \sim \chi_b^2/b$, where $b = (L-1)/\sum_{s=1}^{S} {w_s}^2$.

The approximate t-corrected distribution for Q

From $Q|\, B_\infty, \overline{Q_L}, B_L \sim N(\overline{Q_L}, (1 + L^{-1})B_\infty)$ and $B_L/B_\infty|\, B_L \sim \chi_b^2/b$, we have

$$Q|\,\overline{Q_L}\,, B_L \sim t_b(\overline{Q_L}\,, (1+L^{-1})B_L)$$

(Gelman *et al*. 2004), where $b = (L-1)/\sum_{s=1}^{S} {w_s}^2$ and $w_s = \frac{\frac{1}{B_L^{(s)}}}{\sum_{s=1}^{S}\frac{1}{B_L^{(s)}}}$. Again replacing $\overline{Q_L^{(s)}}$ and $B_L^{(s)}$with the sample estimates $\overline{q_L^{(s)}}$ and $b_L^{(s)}$ yields the desired result.

2.4.3 Randomization Validity

2.4.3.1 Unbiasedness of the Combined Estimator

Under *assumptions 1-3*, the estimates from the synthetic populations of each survey, $\overline{Q_l^{(s)}}$, $s = 1, \dots, S$, are unbiased with respect to repeated sampling from the fixed population (Raghunathan et al. 2003), i.e., $\mathrm{E}(\overline{Q_l^{(s)}}|\,\mathcal{P}_l^{(s)}) = Q$, where $\mathcal{P}_l^{(s)}$ denotes the l^{th} synthetic population for the s[th] survey. Thus,

$$E\left(\overline{Q_\infty} = \sum_{s=1}^{S}\frac{\overline{Q_\infty^{(s)}}}{B_\infty^{(s)}} \Big/ \sum_{s=1}^{S}\frac{1}{B_\infty^{(s)}}\,\middle|\,\mathcal{P}_l^{(s)}, l = 1, \dots, L, \qquad s = 1, \dots, S\right)$$

$$= \sum_{s=1}^{S}\frac{\mathrm{E}(\overline{Q_\infty^{(s)}}|\,\mathcal{P}_l^{(s)})}{B_\infty^{(s)}} \Big/ \sum_{s=1}^{S}\frac{1}{B_\infty^{(s)}}$$

$$= \sum_{s=1}^{S}\frac{Q}{B_\infty^{(s)}} \Big/ \sum_{s=1}^{S}\frac{1}{B_\infty^{(s)}}$$

$= Q$,

which implies that the combined estimator across S surveys is unbiased for the population true value.

2.4.3.2 Gains in Precision

If the synthetic populations are generated properly, $b_{\infty}^{(s)}$, $s = 1, \dots, S$, will be close to or slightly bigger than the variance estimate from the actual data (*Assumption 2 and 3*) because of the information loss when generating the synthetic populations.

Assume the minimum variance estimate among the S surveys is $b_{\infty}^{(1)}$. Then the variance estimates from the S surveys can be written as $b_{\infty}^{(s)} = k^{(s)} * b_{\infty}^{(1)}$ with $k^{(s)} \geq 1, s = 1, \dots, S$. Then the variance estimate of the combined estimator is $b_{\infty} = 1 / \sum_{s=1}^{S} \frac{1}{b_{\infty}^{(s)}} = \frac{1}{\sum_{s=1}^{S} \frac{1}{k^{(s)} * b_{\infty}^{(1)}}} = \frac{1}{\sum_{s=1}^{S} \frac{1}{k^{(s)}}} b_{\infty}^{(1)} < b_{\infty}^{(1)} \leq b_{\infty}^{(s)}$, $s = 1, \dots, S$. The largest gain in precision happens when the variance estimates from the S surveys are equal, i.e., $k^{(s)} = 1, s = 1, \dots, S$. In this situation, the variance of the combined estimator is $1/S$ of the ones from individual surveys. Even though we may lose information when generating the synthetic populations, the combined estimator should still be more precise than those from individual surveys.

2.5 Simulation Study

In the next two sections, we describe two studies to demonstrate the application of the proposed method. There are two purposes of the studies. The first purpose is to evaluate the model-based method to generate synthetic populations that adjusts for the complex sampling design features. The second purpose is to compare the combined estimates with the estimates from individual surveys. In Section 2.5, we conduct a simulation study that involves a population with four normally distributed variables. We use a linear model to impute the unobserved population. In Section 2.6, we evaluate the new approach in a more realistic situation, in which we combine the 2006 National

Health Interview Survey (NHIS) and the 2006 Medical Expenditure Panel Survey (MEPS) to make inference on people's health insurance coverage.

We create a population with strata and clusters within each stratum from the following linear model. The estimand of primary interest is the population mean of Y, $\bar{Y}$.

$Y_{ijk} = 500 + 7Z_{1ijk} + 5Z_{2ijk} + 9Z_{3ijk} + 4.5 * i + u_{ij} + e_{ijk}$,
where, $Z_1 \sim N(3, sd = 0.5)$, $Z_2 \sim N(8, sd = 0.75)$, $Z_3 \sim N(10, sd = 1)$ are the design variables known for the entire population,
$i = 1, \dots, 150$,
$u_{ij} \sim N(0, 0.1), j = 1, \dots, a_i$,
$a_i \sim uniform(2, 52)$ is the number of clusters within stratum i,
$e_{ijk} \sim N(0, 1), k = 1, \dots, b_{ij}$,
$b_{ij} \sim uniform(20, 120)$ is the number of units within cluster j of stratum i.

The population for the simulation study has 240,785 subjects, denoted by $(Y, Z) = (Y, (Z_1, Z_2, Z_3))$. We draw two samples from the population to simulate the data obtained from two surveys: one is drawn using simple random sampling (SRS) and the other stratified clustering sampling with unequal probabilities of selection. The sample size of the simple random sample is 100,000. For the complex sample, we select two clusters from each stratum with probabilities proportional to cluster size (PPS). Within each selected cluster, we select 1/10 of the population. Thus, the probability that unit ijk is selected is

$\Pr(\text{cluster ij is selected}) * \Pr(\text{unit ijk is selected} | \text{cluster ij is selected}) \propto b_{ij}$.

The weights of the sample are calculated by inversing the selection probabilities. Since the number of clusters and units are random, the complex sample sizes are slightly different across replications, which is approximately 2,000. We denote the samples by (Y_{obs}, Z_{obs}) and the unobserved population by (Y_{nob}, Z_{nob}).

For each sample, $L = 100$ synthetic populations are created using the proposed method:

1. Estimate the approximated posterior distribution of the regression coefficients, $N(\hat{\beta}, cov(\hat{\beta}))$,under the linear model, where $\hat{\beta}$ is the point estimate of the regression coefficients obtained from the sample (Y_{obs}, Z_{obs}) after adjusting for the sampling design.
2. Make 100 draws from the posterior distribution, $\beta^{(l)}, l = 1, \ldots, 100$, where $\beta^{(l)} = (\beta_0^{(l)}, \beta_1^{(l)}, \beta_2^{(l)}, \beta_3^{(l)})$.
3. For $\beta^{(l)}, l = 1, \ldots, 100$, impute the unobserved population using the underlying true model: $Y_{nob}^{(l)} = Z_{nob}(\beta^{(l)})^T$ and generate one synthetic population of Y, $(Y_{obs}, Y_{nob}^{(l)})$.

The population mean of Y is estimated from the synthetic populations and the estimates are combined first within surveys using the combining rule developed by Raghunathan *et al.* (2003) and then across two surveys using the combining rule developed in the chapter. We repeat the process 200 times. Specifically, we draw 200 simple random samples and 200 complex samples from the population. Each pair of simple random sample and complex sample is considered to be the observed data from two surveys.

We first evaluate the proposed synthetic population generation method by comparing the following four statistics, the average of the 200 actual sample estimates of $\bar{Y}$, the average of the 200 actual sample standard error estimates (given in the parentheses in Table 2.2), the standard deviation of the 200 actual sample estimates of $\bar{Y}$ (given in the brackets in Table 2.2) and the rate the 95% confidence interval covers population true value. The results are summarized in Table 2.2.

Table 2.2 Estimates from Population, Sample and Synthetic Populations

	Survey 1: SRS		Survey 2: Stratified Clustering Sampling		Population True Value ($\bar{Y}$)
	Sample	Syn.pop (100)	Sample	Syn.pop (100)	
Mean of Point Estimates	972.279	972.281	972.341	972.402	972.343
Mean of SE Estimates	(0.611)	(0.606)	(0.297)	(0.307)	
SD of Point Estimates	[0.608]	[0.610]	[0.286]	[0.304]	
95% CI Coverage rate	95%	95%	96%	95%	

We see that when the underlying true model is used to impute the unobserved values, the synthetic populations preserve the point estimates and variance estimates very well for both the simple random sample and the complex sample. And the loss of information is trivial. Also, the 95% confidence interval coverage rates between the actual data and the synthetic populations are almost identical. This implies the approximate model-based method adjusts for the complex sampling design features.

In the combining survey context, for each replication, we produce the combined point estimate $\bar{y}$ and variance estimate using the combining rule developed in Section 2.4. And then we compare the coverage rate of the 95% confidence interval as well as the empirical mean square error, $eMSE = \sum_{d=1}^{200}(\bar{y}_d - \bar{Y})^2 /200$, where $\bar{y}_d$ is the estimate for replication d $(d = 1, \dots, 200)$. The results are summarized in Table 2.3.

Table 2.3 Individual Survey Estimates and the Combined Estimate

	Survey 1: SRS	Survey 2: Stratified Clustering Sampling	Combined Estimate
Point Est.	972.279	972.341	972.366
SE	0.611	0.297	0.265
95% CI	95%	96%	95%
eMSE	0.378	0.089	0.076

We notice that while the combined estimate has as good 95% CI coverage rate as the estimates from individual surveys and that it has a smaller empirical mean square error than the estimates from both the simple random sample and the complex sample. The gain in precision over the estimate from the simple random sample is very big. This implies the proposed method uses the information from both samples and produces a more accurate and precise estimator.

2.6 Application

In Section 2.5, we use the true model that generates the target population to impute the synthetic populations. Thus, the inference from the synthetic populations is under the best scenario where the imputer's assumed model is also the correct model.

In realistic situations, the exact model that generates the population is not known, and the model of interest may not be linear. To evaluate the proposed combining survey method in a more realistic setting, we use the 2006 NHIS and MEPS data. The goal is to estimate the coverage rate of health insurance for the whole US population and some subdomains. There are three types of health insurance status, covered by any private insurance, covered by government insurance and uninsured. We choose six demographic variables as independent variables: gender, race, census region, education level, age (categorical), and income level (categorical). This gives us a 7-dimensional table with 16,128 cells. The subdomains are created by one demographic variable or the combination of 2 or 3 demographic variables.

Both the 2006 NHIS and MEPS data are multistage probability sample that incorporates stratification, clustering and oversampling of some subpopulations (e.g.,

Black, Hispanic, and Asian). We delete the cases with item missing values and focus on our simulation on the complete cases. This results in 20,147 and 20,893 cases in the NHIS and MEPS respectively. We recode the variables into the same categories. The coding of the variables is shown in Table 2.4 below.

Table 2.4 Variables and Response Categories for the 2006 NHIS and MEPS

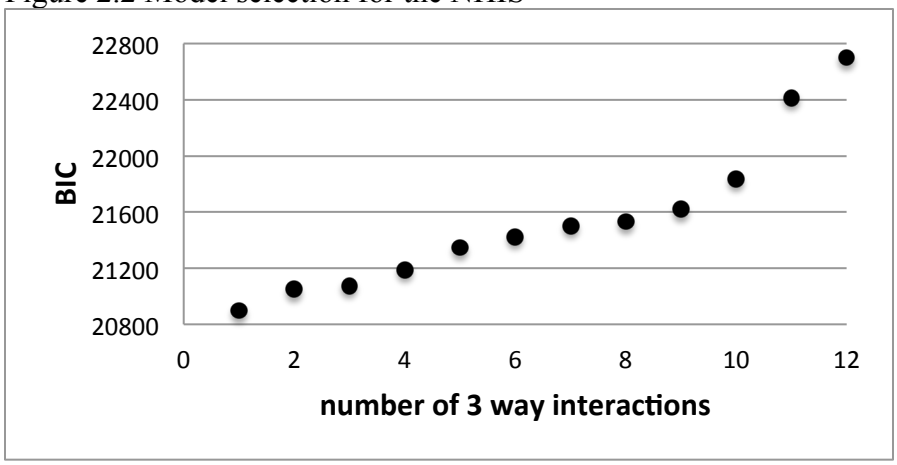
Variables of Interest	Response Categories
Age	1: [18,24] 2: [25,34] 3: [35,44] 4: [45,54] 5: [55,64] 6: >=65
Census Region	1: Northeast 2: Midwest 3: South 4: West
Education	1: Less than high school 2: High school 3: Some college 4: College
Gender	1: Male 2: Female
Health Insurance Coverage	1: Any Private Insurance 2: Public Insurance 3: Uninsured
Income	1: (0,10000) 2: [10000,15000) 3: [15000,20000) 4: [20000,25000) 5: [25000,35000) 6: [35000,75000) 7: >=75000
Race	1: Hispanic 2: Non-Hispanic White 3: Non-Hispanic Black 4: Non-Hispanic All other race groups

If the model that generates the synthetic populations fails to include important terms, we could lose a fairly large amount of information. In the other hand, if we include too many terms, we may have highly unstable estimates of the log-linear model coefficients resulting in spurious variability in the synthetic populations. So we need to determine the model that fits the data best.

We use a forward model selection approach to decide the level of interaction that should be included into the model. The Bayesian information criterion (BIC) is used to

compare different models. Specifically, we start the model with all 7 main effects and all 2 way interactions. We add one 3 way interaction at a time, choose the most significant 3 way interactions and then calculate the BIC of this model. Then we add the other 3 way interactions into the current model (the one with the most significant 3 way interaction) one at a time, choose the most significant one and calculate the BIC. We repeat this until the BIC starts to increase. Figure 2.2 below is the BIC versus the number of 3 way interactions for the NHIS. As we can see, the BIC is increasing since the first 3 way interaction is added, which suggests the model with all main effects and 2 way interactions is sufficient. Also, the Pearson Chi-square for this model is close to the number of degrees of freedom, which is also a sign of goodness of fit. We get the same model for the MEPS data following the same model selection procedure.

Figure 2.2 Model selection for the NHIS



Following the proposed method, we generate 100 synthetic populations for each survey. We analyze them as simple random samples and combine the estimates from the synthetic populations within each survey using the combining rules for synthetic data. The results are summarized in Table 2.5.

Table 2.5 Estimates from Actual Data and from Synthetic Populations for the 2006 NHIS and MEPS

Domain	Actual Data (Complex Design)			Synthetic Populations	
	Types	NHIS	MEPS	NHIS	MEPS
Whole Population			Proportion		
	Private	0.746	0.735	0.7457	0.734
	Public	0.075	0.133	0.0757	0.133
	Uninsured	0.179	0.132	0.1785	0.132
			Variance		
	Private	2.46E-05	2.78E-05	2.66E-05	2.86E-05
	Public	6.29E-06	1.44E-05	7.99E-06	1.77E-05
	Uninsured	1.84E-05	1.41E-05	1.81E-05	1.56E-05
Male			Proportion		
	Private	0.740	0.735	0.7397	0.735
	Public	0.060	0.101	0.060	0.102
	Uninsured	0.200	0.164	0.2000	0.164
			Variance		
	Private	3.32E-05	3.87E-05	3.70E-05	3.52E-05
	Public	6.82E-06	1.53E-05	7.91E-06	1.91E-05
	Uninsured	2.94E-05	2.64E-05	3.19E-05	2.56E-05
Hispanic			Proportion		
	Private	0.494	0.506	0.4933	0.506
	Public	0.096	0.161	0.0969	0.161
	Uninsured	0.410	0.333	0.4099	0.333
			Variance		
	Private	1.24E-04	1.73E-04	1.33E-04	2.08E-04
	Public	2.57E-05	8.03E-05	3.28E-05	9.46E-05
	Uninsured	1.23E-04	1.19E-04	1.32E-04	1.67E-04
Non-Hispanic White	Proportion				
	Private	0.805	0.788	0.8045	0.788
	Public	0.062	0.116	0.062	0.117
	Uninsured	0.134	0.096	0.1337	0.096
			Variance		
	Private	2.99E-05	3.35E-05	3.07E-05	3.98E-05
	Public	8.20E-06	1.81E-05	1.10E-05	2.45E-05
	Uninsured	2.02E-05	1.51E-05	1.82E-05	1.82E-05
Non-Hispanic White & Income [25,000, 35,000)			Proportion		
	Private	0.827	0.813	0.8404	0.838
	Public	0.039	0.079	0.0371	0.067
	Uninsured	0.134	0.108	0.1225	0.096
			Variance		
	Private	1.00E-04	1.39E-04	6.80E-05	8.59E-05
	Public	2.82E-05	6.31E-05	1.79E-05	4.25E-05
	Uninsured	7.24E-05	8.92E-05	4.38E-05	5.79E-05

We see the point estimates and the variance estimates from the synthetic data are similar to those from the actual data after complex sampling design features are taken into account. However, the correspondence is less exact for the smaller subdomains such as the Non-Hispanic white people with Income between 25,000 and 35,000 per year. The reason may be that the imputation model that is fitted to the whole sample globally may not hold well for this small domain of size 2,193. Next, we produce the combined estimates using the combining rules for multiple surveys. The results are summarized in Table 2.6. From the table, we notice the variance estimates for the combined estimator are much smaller than the ones from individual surveys. For example, the combined estimator is 82% more precise than the estimates from the NHIS and 256% more precise than the estimates from the MEPS on average. The largest increase in precision over the NHIS is by 191% for estimating the proportion of Non-Hispanic white people with Income between 25,000 and 35,000 per year who are uninsured and the largest increase in precision over the MEPS is by 266% for estimating the proportion of Non-Hispanic white people with Income between 25,000 and 35,000 per year 266% who are covered by any private insurance.

Table 2.6 Estimates from Individual Surveys and the Combined Estimates for the 2006 NHIS and MEPS

Domain	Actual Data (Complex Design)			Combined Estimates
	Types	NHIS	MEPS	
Whole Population			Proportion	
	Private	0.746	0.7348	0.740
	Public	0.075	0.1330	0.094
	Uninsured	0.179	0.1322	0.154
			Variance	
	Private	2.46E-05	2.78E-05	1.38E-05
	Public	6.29E-06	1.44E-05	5.50E-06
	Uninsured	1.84E-05	1.41E-05	8.38E-06
Male			Proportion	
	Private	0.740	0.7354	0.737
	Public	0.060	0.1010	0.072
	Uninsured	0.200	0.1636	0.180
			Variance	
	Private	3.32E-05	3.87E-05	1.80E-05
	Public	6.82E-06	1.53E-05	5.59E-06
	Uninsured	2.94E-05	2.64E-05	1.42E-05
Hispanic			Proportion	
	Private	0.494	0.5057	0.498
	Public	0.096	0.1608	0.113
	Uninsured	0.410	0.3335	0.376
			Variance	
	Private	1.24E-04	1.73E-04	8.11E-05
	Public	2.57E-05	8.03E-05	2.44E-05
	Uninsured	1.23E-04	1.19E-04	7.37E-05
Non-Hispanic White			Proportion	
	Private	0.805	0.7877	0.797
	Public	0.062	0.1161	0.079
	Uninsured	0.134	0.0962	0.115
			Variance	
	Private	2.99E-05	3.35E-05	1.73E-05
	Public	8.20E-06	1.81E-05	7.59E-06
	Uninsured	2.02E-05	1.51E-05	9.10E-06
Non-Hispanic White & Income [25,000, 35,000)			Proportion	
	Private	0.827	0.8132	0.839
	Public	0.039	0.0792	0.046
	Uninsured	0.134	0.1076	0.111
			Variance	
	Private	1.00E-04	1.39E-04	3.80E-05
	Public	2.82E-05	6.31E-05	1.26E-05
	Uninsured	7.24E-05	8.92E-05	2.49E-05

2.7 Discussion

In this chapter, we propose a new method to combine information from multiple complex surveys. We evaluate the new method by using a simulation study and applying it to combine information about health insurance status from the 2006 NHIS and MEPS. Both show the combined estimate is more precise compared to the estimates from individual surveys. The simulation study uses the underlying true model to generate synthetic populations while adjusting for the sampling designs. We have no information loss in the sense that the sampling properties of inferences from the synthetic population and the actual sample are very similar. Then we combine the estimates from two samples and the combined estimate outperforms the estimates from individual surveys with respect to mean square error while retaining correct 95% confidence interval coverage. In the application, although there is some loss of information due to the imputation model is not the underlying true model, the combined estimates of health insurance status that use the information from two surveys are still more precise than the ones from individual surveys.

The quality of inferences of the proposed method clearly depends on the imputation models. It is possible to obtain valid inferences from combining multiple surveys if relationship is accurately modeled in the imputation models. On the other hand, when the imputation model is misspecified, the inference from the synthetic populations may not be valid, which implies the combined estimates may not be valid. For example, in the second simulation study where we combined the 2006 NHIS and MEPS, the inference from the synthetic populations does not simulate that of the actual data well for

the smallest domain. When developing an accurate imputation model is impossible due to small sample size or complicated data structure, we could use nonparametric methods to protect against model misspecification. This will be discussed in Chapter 3.

This new combining survey method has two major advantages over the existing methods. First, it adjusts for the complex sampling design features when imputing the unobserved population. Since the synthetic populations can be analyzed as simple random samples, information from other surveys can be used to adjust for the nonsampling errors and/or filling in the missing variables. For example, one of the greatest interests in combining survey area is in the situation that each survey only covers a subset of variables of interest and we have to combine multiple surveys of different sampling designs to obtain all the variables of interest.

Another advantage of this method is it has no limitation on the number of surveys to be combined as long as the surveys have the same underlying population. The proposed method that adjusts for the complex sampling design features can be applied to each survey independently. After the missing information is imputed, regardless the number of surveys to be combined, we just need to combine the estimates from each survey using the combing rule developed in this chapter. It would be interesting to see how much more gains in precision we could obtain when we combine more than two surveys. While this chapter aims at laying down the theoretical foundation, we will extend and evaluate the new method in more general situations in the future research.

CHAPTER 3

A NONPARAMETRIC METHOD TO GENERATE SYNTHETIC POPULATIONS TO ADJUST FOR THE COMPLEX SAMPLING DESIGN FEATURES

Outside of the survey sampling literature, samples are often assumed to be generated by simple random sampling process that produces independent and identically distributed (IID) samples. Many statistical methods are developed largely in this IID world. Application of these methods to data from complex sample surveys without making allowance for the survey design features can lead to erroneous inferences. Hence, much time and effort have been devoted to develop the statistical methods to analyze complex survey data and account for the sample design. An alternative to tailor the methods to fit the data is to work backwards, tailoring the data to fit the methods. The first method developed along these lines is the inverse sampling algorithm (Hinkins, Oh and Scheuren, 1997). In this chapter, we propose a new nonparametric method to invert the complex sampling design features and generate simple random samples from a missing data point

of view. This method achieves the same goal as the inverse sampling does, making adjustment on the complex data so that they can be analyzed as simple random samples. We apply the method to two sample designs, one-stage stratified sampling and stratified clustering sampling. Both situations use weighting to adjust for the unequal selection probabilities. We use the nonparametric method to generate synthetic populations for the 2006 National Health Interview Survey (NHIS), the Behavioral Risk Factor Surveillance System (BRFSS) and the Medical Expenditure Panel Survey (MEPS). We then apply this method in the new combining survey framework developed in Chapter 2 and produce the combined estimates of the health insurance coverage rates for the US population.

3.1 Introduction

The development of survey sampling techniques is an extraordinary achievement (Hansen 1987, Kish 1995). The richness in modern sampling techniques may isolate the analysis of survey data from the classical statistics, which has mainly been developed for simple random samples or more recently, one-stage cluster samples without concerning for issues such as stratification, unequal probability of selection, nonresponse bias or calibration. Major efforts of modern survey statistics focus on developing methods that are appropriate to analyze complex survey data (Skinner, Holt and Smith 1989). Hinkins, Oh and Scheuren (1997) proposed an inverse sampling design algorithm that connects the survey statistics and the classical statistics from another perspective. Instead of developing new statistical techniques to fit the data, the inverse sampling technique resample from the data to produce equivalent simple random samples that can be analyzed using the classical statistical methods. Adapting a quote from Hinkins, Oh and Scheuren (1997): “If you only have a hammer, every problem turns into a nail!”. Their

basic idea is to choose a subsample that has a simple random sample structure unconditionally. The subsample is often much smaller than the original sample so they propose to repeat the process independently many times and average the results to increase the precision. They also described the exact or approximate inverse sampling schemes under multiple situations such as the stratified simple random sampling, one-stage cluster sampling and two-stage cluster sampling. However, this new idea is not used widely in practice mainly because it is extremely computionally intensive and the precision losses are often substantial.

In the last chapter, we proposed a new method from a missing data perspective for the purpose of combining multiple surveys. Unlike the inverse sampling technique that assumes the population is no longer available and we can only draw the subsamples from the original sample, the new method assumes the sample is drawn from a finite population which can be recovered after we impute the unobserved part of it. We developed the imputation model from a Bayesian framework. Specifically, we approximate the posterior distribution of the model parameters by the asymptotic normal distribution. The mean and covariance matrix of the normal distribution are estimated after complex sampling design features are taken into account.

However, all statistical models are simplifications and hence subject to some degree of misspecification (Little 2004). The major weakness of a model-based method is if the model is seriously misspecified, it may yield invalid inferences (Little 2004). Model misspecification includes neglecting to include an important covariate, misspecifiying its functional form, or making an erroneous distributional assumption. Although the general steps to apply the model-based method are the same across

situations, the details could vary greatly in practice. First, we need to consider the relationships among the variables of interest and determine an appropriate model that fits the data, which may be hard if the data contains different types of variables. After we determine the model, we also we need to develop specific strategies for both model selection and model fitting. This is more challenging when the data that is obtained using different complex sampling designs.

In this chapter we propose a nonparametric method as a counterpart of the model-based method to generate synthetic populations. The nonparametric method focuses on the design of the survey so we can avoid modeling the complicated relationships among the variables in the data. The basic idea is to resample from the actual data to impute the unobserved part of the population. Bayesian bootstrap methods are used in this process. Since it achieves the same goal of the inverse sampling technique, it can be treated as the Bayesian finite population version of inverse sampling.

This chapter is organized as follows: Section 3.2 reviews and summarizes different bootstrap methods. Section 3.3 presents the proposed method under two situations, one-stage stratified sampling and stratified clustering sampling. Both situations also have samples obtained with unequal selection probabilities. Section 3.4 proves that the point estimate from the synthetic populations is unbiased for the population true value and that the variance estimate from the synthetic populations is approximately unbiased for the one that is obtained from actual data after adjusting for the complex sampling design features. Section 3.5 provides a simulation study to evaluate the performance of the nonparametric method. Section 3.6 applies the method to estimate health insurance coverage rates using the 2006 NHIS, MEPS and BRFSS data. We also

applied the combining survey method proposed in Chapter 2 and produced the combined estimates of people's health insurance coverage rates. Concluding remarks are provided in Section 3.7.

3.2 Background

3.2.1 The Bootstrap

The bootstrap method is first proposed by Efron (1979) in the case of an independent and identically distributed sample. It has great applications in statistics for situations where explicit formulae for measuring variances and conducting significance tests are intractable. The bootstrap draws multiple simple random samples with replacement from the original sample to simulate the sampling distribution of a statistic of interest. It essentially assumes the sample cumulative distribution function (cdf) of the statistic is the population cdf.

Rao and Wu (1988) extend Efron's bootstrap method to complex survey data, especially those obtained from stratified clustering sampling. Suppose a complex sample contains H strata and there are c_h clusters within stratum $h, h = 1, \dots, H$. Denote C as the total number of clusters in the data, i.e., $C = \sum_{h=1}^{H} c_h$. Suppose the statistic of interest is Q. The bootstrap method is established in the following steps (Rao and Wu 1988).

1. In stratum $h, h = 1, \dots, H$, draw a simple random sample with replacement (SRSWR) of m_h from the c_h clusters. Let m_{hi}^* denote the number of times that cluster $i, i = 1, \dots, c_h$ is selected from stratum h, so that

$$m_h = \sum_{i=1}^{c_h} m_{hi}^*.$$

For each element $k, k = 1, \dots, N_{hi}$ within cluster i from stratum h, we denote its original weight by w_{hik}. Then we create the replicate weight as:

$$w^*_{hik} = w_{hik}\left(\left(1 - \sqrt{\frac{m_h}{c_h - 1}}\right) + \sqrt{\frac{m_h}{c_h - 1}}\frac{c_h}{m_h}\, m^*_{hi}\right).$$

To ensure all the replicate weights are non-negative , $m_h \leq (c_h - 1)$.

2. Suppose we generate B bootstrap samples. For each bootstrap sample $b, b = 1, \dots, B$, calculate the estimate of the statistic of interest Q using the replicate weights ${w^*_{hik}}^{(b)}$, denoted by $\widetilde{Q^{*(b)}}$. Similarly to Efron's bootstrap, $\widetilde{Q^{*(b)}}$, $b = 1, \dots, B$, simulate the sampling distribution of Q. The point estimate of Q is obtained from

$$\overline{Q_{boot}} = \sum_{b=1}^{B} \widetilde{Q^{*(b)}}\, /B.$$

The variance of $\overline{Q_{boot}}$ is calculated from

$$var_{boot}(\overline{Q_{boot}}) = \frac{1}{B}\sum_{b=1}^{B}(\widetilde{Q^{*(b)}} - \overline{Q_{boot}})^2,$$

which reflects the change in variance caused by stratification and clustering. A special case is when there are 2 PSUs in each stratum. In this setting, the only choice for the value of m_h is $m_h = 1$.

3.2.2 The Bayesian Bootstrap

The Bayesian Bootstrap is developed by Rubin (1981) as a Bayesian analogue of the bootstrap. It is quite similar to the bootstrap operationally and inferentially. For example, Lo (1987) showed that the Bayesian bootstrap has the same desirable large

sample properties as Efron's bootstrap. But the Bayesian bootstrap performs better for small samples because of its Bayesian justification. In the other hand, the Bayesian bootstrap and bootstrap have different interpretations. The bootstrap simulates the sampling distribution of a statistic estimating the parameter, while the Bayesian bootstrap simulates the posterior distribution of the parameters of interest. Based on this posterior distribution, we can obtain the posterior predictive distribution of the unobserved subjects given the sample, from which the unobserved subjects of the population can be drawn.

The Bayesian bootstrap is established by making draws from a posterior distribution of the parameters that is obtained from a Dirichlet prior and a multinomial likelihood. It is first developed for simple random sampling with replacement. For example, assume the variable of our interest for the population is Y and a sample of size n is denoted by $(y_1, \dots, y_n)$. We will see that the Bayesian bootstrap draws the subjects in the sample and thus is not variable-specific. Once a subject is selected, all the variables are selected. So $y_i, i = 1, \dots, n$ actually denote the n subjects in the sample. Operationally, each BB sample is selected by the following two steps (Rubin 1981).

1. Draw n uniform random numbers between 0 and 1, and let their ordered values be $a_1, \dots, a_n$ and also let $a_0 = 0$ and $a_n = 1$.
2. Draw each of the n values in the BB sample by drawing from $(y_1, \dots, y_n)$ with probabilities $((a_1 - a_0), (a_2 - a_1), \dots, (1 - a_{n-1}))$.
3. Suppose we generate B BB samples. Then the B BB replications gives the Bayesian bootstrap distribution of Y (or posterior predictive distribution of unobserved Y) and thus of any parameter of this distribution. For example, for

each BB sample $b, b = 1, \dots, B$, we calculate the estimate of the statistic of interest $Q(Y)$, denoted by $\widetilde{Q^{*(b)}}$. Then the point estimate of Q is obtained from

$$\overline{Q_{BB}} = \sum_{b=1}^{B} \widetilde{Q^{*(b)}} / B.$$

The variance of $\overline{Q_{BB}}$ is calculated from

$$var_{BB}(\overline{Q_{BB}}) = \frac{1}{B}\sum_{b=1}^{B}(\widetilde{Q^{*(b)}} - \overline{Q_{BB}})^2.$$

Consider the similarity between the bootstrap and Bayesian bootstrap, the rationale behind how the bootstrap adjusts for complex sampling design features can be naturally generalized to the Bayesian bootstrap. However, unlike the bootstrap replications simulating the sampling distribution of the statistic of interest, the BB samples simulate the posterior distribution of the statistic. We can use the same scheme to calculate the replicate weights for the BB samples.

3.2.3 Finite Population Bayesian Bootstrap

The finite population bootstrap (FPB) was first proposed by Gross (1980). Bickel and Freedman (1984) and Chao and Lo (1985) provided a first-order asymptotic justification for the FPB mean. This method assumes $(y_1, \dots, y_n)$ is a simple random sample from a finite population $(Y_1, \dots, Y_N)$ and the population size N is an integer multiple of the sample size n, that is, $N = kn$. Then, FPB replicates the sample k times to create the FPB population. Each FPB sample is drawn by simple random sampling without replacement from the FPB population to obtain $(y_1^*, \dots, y_n^*)$. The FPB is developed from a frequentist's point of view and is equivalent to Efron's bootstrap for a large population.

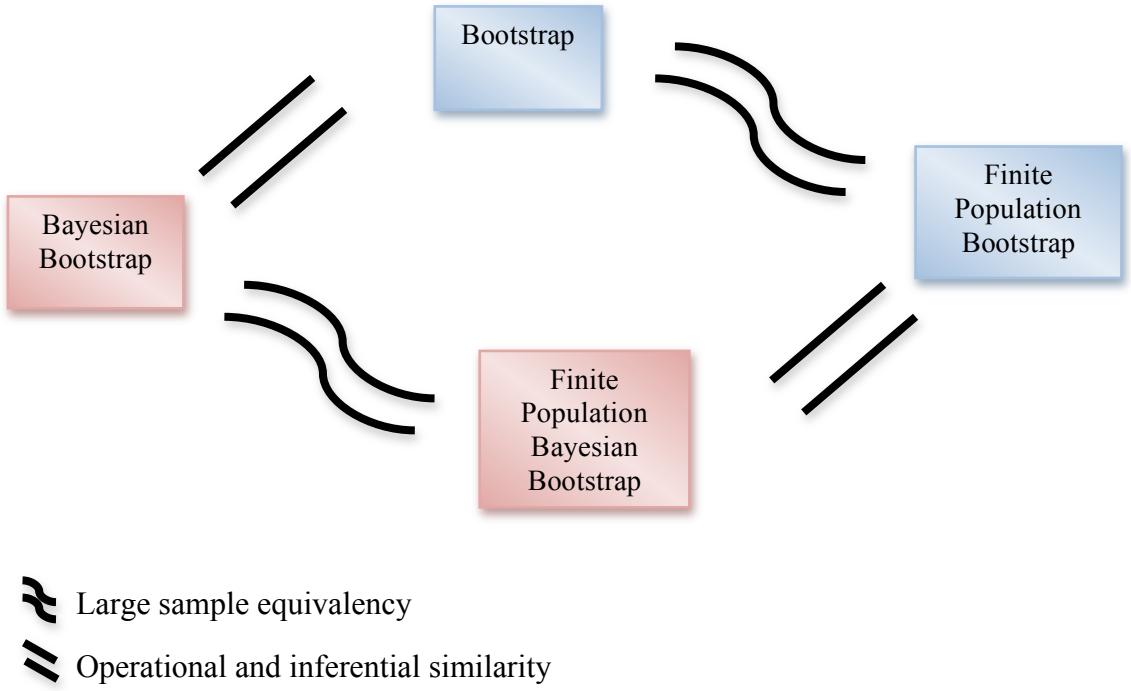
Lo (1986, 1988) developed the finite population Bayesian bootstrap (FPBB) as a Bayesian analogue of FPB. He simulated a posterior distribution with respect to a "flat" Dirichlet-multinomial prior (Ferguson 1973). However, the FPBB is extremely computationally intensive to make draws from the posterior predictive distribution since it involves the calculation of the gamma functions. This further requires the calculation of the number of possibilities that we choose n units out of N and n and N are usually huge for survey data. To improve its practicability, we make draws from the FPBB posterior predictive distribution using a "Pólya urn scheme" procedure (Lo 1988). Suppose an urn contains a finite number of balls. A Pólya sample of size m is selected by first selecting a ball at random from the urn and returning the selected ball into the urn, then putting one same ball into the urn and repeating this process until m balls have been selected. Each replication of the FPBB is drawn using the following steps:

Step 1. Draw a Pólya sample of size $N - n$, denoted by $(y_1^*, \dots, y_{N-n}^*)$ from the urn $\{y_1, \dots, y_n\}$.

Step 2. Form the FPBB population $y_1, \dots, y_n, y_1^*, \dots, y_{N-n}^*$.

It appears that the FPBB, FPB, Bayesian bootstrap and Efron's bootsrap are closely related (Figure 3.1). For example, it is shown that the FPBB and FPB share similar operational characteristics and small sample properties. The FPBB reduces to the Bayesian bootstrap for a large population (Lo 1988). Lo also provided a first-order asymptotic equivalence of the FPBB and FPB in his 1988 paper.

Figure 3.1 Equivalence of bootstrap, Bayesian bootstrap, finite population bootstrap and finite population Bayesian bootstrap



3.3 Nonparametric Method to Generate Synthetic Populations

In this section, we present the nonparametric method to generate synthetic populations that adjusts for the complex sampling design features. The idea is to treat the unobserved part of the population as missing data and impute it by making draws from the actual data. Once we have a draw from the posterior predictive distribution of the whole population, the complex sampling design features will be of no use and we can analyze it as a simple random sample.

Cohen (1997) extended the FPBB procedure to adjust for the unequal probabilities of selection. Assume $(y_1, \dots, y_n)$ is a sample from a finite population $(Y_1, \dots, Y_N)$ with weights $(w_1, \dots, w_n)$. The procedure has two steps:

Step 1. Draw a sample of size $N-n$, denoted by $(y_1^*, \ldots, y_{N-n}^*)$, as follows:

determine y_k^* by drawing from $(y_1, \ldots, y_n)$ in such a way that y_i is selected with probability $\frac{w_i - 1 + l_{i,k-1} * (N-n)/n}{N-n+(k-1)*(N-n)/n}$, where w_i is the weight of unit i and $l_{i,k-1}$ is the number of bootstrap selections of y_i among $y_1^*, \ldots, y_{k-1}^*$.

Step 2. Form the FPBB population $y_1, \ldots, y_n, y_1^*, \ldots, y_{N-n}^*$.

Cohen (1997) provided neither theoretical proof nor empirical research to evaluate this procedure. Theorem 1 and its proof below provide theoretical justification for FPBB Polya urn scheme.

Theorem 3.1 Assume $(y_1, \ldots, y_n)$ is a sample from a finite population $(Y_1, \ldots, Y_N)$ drawn with unequal probabilities, and the weights of the sample are normalized to the population size, *i.e.*, $\sum_{i=1}^{n} w_i = N$. Then, FPBB Polya urn scheme results in the same draws for the unobserved part of the population as the values drawn from the posterior predictive distribution obtained from FPBB.

Proof:

The idea is to prove the posterior predictive distribution of the unobserved values given the observed values obtained from FPBB is the same as the posterior predictive distribution obtained from the FPBB Polya urn scheme.

Finite Population Bayesian Bootstrap

Assume the observed data has n unique units, which are selected with unequal probabilities and the weights of the sample are normalized to the population size, *i.e.*, $\sum_{i=1}^{n} w_i = N$. For any variable of interest Y, denote the observed values by $(y_1, \ldots, y_n)$

and denote the unique values of $(y_1, \dots, y_n)$ by $(\tilde{y}_1, \dots, \tilde{y}_{n^*})$ with $n^* \leq n$. FPBB is constructed by assuming a non-informative Dirichlet prior and a multinomial likelihood for the data, i.e.,

Non-informative Dirichlet prior: $\pi(\tilde{\theta}_1, \dots, \tilde{\theta}_{n^*}) \propto \prod_{i=1}^{n^*} \tilde{\theta}_i^{-1}$

Data: multinomial distribution: $\pi(y_{obs}|\tilde{\theta}_1, \dots, \tilde{\theta}_{n^*}) \propto \prod_{i=1}^{n^*} \tilde{\theta}_i^{\tilde{w}_i}$,

where $\tilde{w}_i = \sum_{j=1}^{n} I(y_j = \tilde{y}_i)\, w_j, \;\; i = 1, \dots, n^*$.

Without loss of generality, in this proof, we assume $(y_1, \dots, y_n)$ are unique, i.e., $y_i = \tilde{y}_i, w_i = \tilde{w}_i, \theta_i = \tilde{\theta}_i, i = 1, \dots, n^* = n$. Then the prior and the likelihood become:

Non-informative Dirichlet prior: $\pi(\theta_1, \dots, \theta_n) \propto \prod_{i=1}^{n} \theta_i^{-1}$

Data: multinomial distribution: $\pi(y_{obs}|\theta) \propto \prod_{i=1}^{n} \theta_i^{w_i}$.

The posterior predictive distribution is given by

$$\Pr(\, r_1 \text{ units of values } y_1, \dots, r_n \text{ units of values } y_n, \sum_{i=1}^{n} r_i = N' \,|y_{obs})$$

$$= \frac{\pi(y_{obs}, y_{nob})}{\pi(y_{obs})} = \frac{\int_0^1 \dots \int_0^1 \pi(y_{obs}, y_{nob}, \theta) d\theta_1 \dots d\theta_n}{\int_0^1 \dots \int_0^1 \pi(y_{obs}, \theta) d\theta_1 \dots d\theta_n}$$

$$= \frac{\int_0^1 \dots \int_0^1 \pi(y_{nob}|y_{obs}, \theta)\pi(y_{obs}|\theta)\pi(\theta) d\theta_1 \dots d\theta_n}{\int_0^1 \dots \int_0^1 \pi(y_{obs}|\theta)\pi(\theta) d\theta_1 \dots d\theta_n}$$

$$= \frac{\int_0^1 \dots \int_0^1 \pi(y_{nob}|\theta)\pi(y_{obs}|\theta)\pi(\theta) d\theta_1 \dots d\theta_n}{\int_0^1 \dots \int_0^1 \pi(y_{obs}|\theta)\pi(\theta) d\theta_1 \dots d\theta_n}$$

$$= \frac{\int_0^1 \dots \int_0^1 \prod_{i=1}^{n} \theta_i^{r_i} \prod_{i=1}^{k} \theta_i^{w_i} \prod_{i=1}^{k} \theta_i^{-1}\, d\theta_1 \dots d\theta_n}{\int_0^1 \dots \int_0^1 \prod_{i=1}^{n} \theta_i^{w_i} \prod_{i=1}^{k} \theta_i^{-1}\, d\theta_1 \dots d\theta_n}$$

$$= \frac{\int_0^1 \dots \int_0^1 \prod_{i=1}^{n} \theta_i^{r_i + w_i - 1}\, d\theta_1 \dots d\theta_n}{\int_0^1 \dots \int_0^1 \prod_{i=1}^{n} \theta_i^{w_i - 1}\, d\theta_1 \dots d\theta_n}$$

$$= \frac{\prod_{i=1}^{n}\left(\frac{\Gamma(r_i + w_i)}{\Gamma(w_i)}\right)}{\Gamma(N + N')/\Gamma(N)}.$$

However, drawing the unobserved units from this posterior predictive distribution is very computationally intensive especially when N' is large.

Polya Urn Scheme

The observed data can be viewed as an urn contains w_1 balls of value y_1, …, w_n balls of value y_n, where $\sum_{i=1}^{n} w_i = N$. The Polya urn scheme draws one ball at random from the urn and then replaces it with a ball with the same value along with an additional ball with the same value. Since the number of the balls of different values is unequal, the selection probability of the ball of value y_i is $\frac{w_i}{N}, i = 1, \ldots, n$. Given the observed data, the probability that we draw N'balls and that the first r_1 balls have value y_1 through the last r_n balls have value y_n is:

$$\frac{w_1 * (w_1 + 1) * \ldots * (w_1 + r_1 - 1)}{N * (N + 1) * \ldots * (N + r_1 - 1)} * \frac{w_2 * (w_2 + 1) * \ldots * (w_2 + r_2 - 1)}{(N + r_1) * (N + r_1 + 1) * \ldots * (N + r_1 + r_2 - 1)}$$

$$* \ldots * \frac{w_n * (w_n + 1) * \ldots * (w_n + r_k - 1)}{(N + \sum_{i=1}^{n-1} r_i) * (N + \sum_{i=1}^{n-1} r_i + 1) * \ldots * (N + \sum_{i=1}^{n} r_i - 1)}$$

$$= \frac{\prod_{i=1}^{n}\left(\frac{\Gamma(r_i + w_i)}{\Gamma(w_i)}\right)}{\Gamma(N + N')/\Gamma(N)}.$$

The probability of selecting any permutation of the N' balls that have r_1 balls of value y_1 through r_n balls of value y_n is the same because the ordering only affects the permutation of the nominators. So for the FPBB Polya urn scheme,

$$\Pr(\, r_1 \text{ units of values } y_1, \ldots, r_n \text{ units of values } y_n, \sum_{i=1}^{n} r_i = N' \,|\, y_{obs})$$

$$= \frac{\prod_{i=1}^{n}\left(\frac{\Gamma(r_i + w_i)}{\Gamma(w_i)}\right)}{\Gamma(N + N')/\Gamma(N)},$$

which is the same as the probability for the FPBB.

Now, the goal is to draw the unobserved part of the population from the sample $(y_1, \ldots, y_n)$, which together with the sample produce the synthetic population. Assume the weights of the sample are normalized to the population size, *i.e.*, $\sum_{i=1}^{n} w_i = N$ and that in the unobserved part of the population, there are $w_i - 1$ balls of value $y_i, i = 1, \ldots, n$, which implies the probability of selecting the ball of value y_i in the sample is $\frac{w_i - 1}{N - n}, i = 1, \ldots, n$. This can be further converted into a Polya urn problem, where in the urn, there are $\frac{w_i - 1}{N - n} n$ balls of value $y_i, i = 1, \ldots, n$. The FPBB Polya urn scheme suggests we draw one ball at random and then replace the selected ball in the urn along with an additional ball of the same value. It is straightforward to show that y_k^* out of the unobserved population $(y_1^*, \ldots, y_{N-n}^*)$ should be selected in such a way that y_i is selected with probability

$$\frac{\frac{w_i - 1}{N - n} n + l_{i,k-1}}{n + (k-1)} = \frac{w_i - 1 + l_{i,k-1} * (N-n)/n}{N - n + (k-1) * (N-n)/n},$$

where w_i is the weight of unit i and $l_{i,k-1}$ is the number of bootstrap selections of y_i among $y_1^*, \ldots, y_{k-1}^*$. Thus, the $N - n$ draws $y_1^*, \ldots, y_{N-n}^*$ along with the original n balls in the urn, $y_1, \ldots, y_n$, produce one synthetic population. This completes our proof.

To adjust for the complex sampling design features, we should apply FPBB Polya urn scheme to adjust for both clustering and unequal probability of selection. For example, for a one-stage stratified sample, we could use FPBB Polya urn scheme to draw the unobserved population from the actual data. Once we have the whole population

imputed, the complex sampling design features can be ignored and we can analyze them as simple random samples.

For a stratified clustering sampling, the idea is to first apply FPBB Polya urn scheme to impute the unobserved clusters within each stratum. Then within each cluster, we apply FPBB Polya urn scheme to draw the unobserved part of the population. For example, suppose a complex sample contains H strata and there are c_h clusters within stratum $h, h = 1, \dots, H$. Denote c as the total number of clusters in the actual data, i.e., $c = \sum_{h=1}^{H} c_h$. We use the capitalized letters to denote the number of clusters in the population, i.e., $C = \sum_{h=1}^{H} C_h$. The first step is to use FPBB Polya urn scheme to impute the unobserved clusters within each stratum, $c_1^*, \dots, c_{C_h - c_h}^*$, which together with the observed clusters provide the clusters in Stratum h in the population. Then within each of the C_h cluster, we apply FPBB Polya urn scheme to impute the unobserved units so that we have the whole population. However, it is very hard in practice to accurately estimate the probabilities of selecting clusters based on the information that survey agencies typically release to public. Thus, we propose the following approximated steps to generate synthetic populations for stratified clustering sampling.

Step 1: Use the Bayesian Bootstrap to adjust for stratification and clustering

Assume the sample is obtained using a stratified clustering sampling with unequal selection probabilities. We first draw a Bayesian bootstrap sample of the clusters within each stratum and then repeat *L* times to produce *L* Bayesian bootstrap (BB) samples denoted by $S_1, \dots, S_L$. Considering the equivalence between the classical bootstrap and Bayesian bootstrap, we calculate the replicate weights for each BB sample as Rao and

Wu (1988) suggested such that each of the Bayesian bootstrap samples has weights, $w^{*(l)} = \left\{w_{hik}^{*(l)},\ h = 1, \dots, H,\ i = 1, \dots, c_h,\ k = 1, \dots, N_{hi}\right\}$, where $l = 1, \dots, L$, $w_{hik}^{*} = w_{hik}\left(\left(1 - \sqrt{\frac{m_h}{c_h - 1}}\right) + \sqrt{\frac{m_h}{c_h - 1}}\frac{c_h}{m_h}\, m_{hi}^{*}\right)$. Assume the weighted estimate of Q for replication l is denoted by $\widetilde{Q_l}, l = 1, \dots, L$, which simulate the posterior distribution of Q. Thus, the average across $\widetilde{Q_l}, l = 1, \dots, L$, provides an unbiased estimate for Q. The between-variance of $\widetilde{Q_l}, l = 1, \dots, L$, is the variance estimate after the complex sampling design features are accounted for.

If the sample is selected using a stratified sampling mechanism with unequal selection probabilities within strata, we apply the Bayesian bootstrap procedure to the subjects within each stratum and calculate the replicate weights as Rao and Wu (1988) suggested.

This step generates *L* Bayesian bootstrap samples which essentially are *L* draws from the posterior predictive distribution of the unobserved clusters given the actual data. However, the units for the *L* Bayesian bootstrap samples still have weights and cannot be analyzed as simple random samples.

Step 2: Use FPBB Polya urn scheme to adjust for weighting

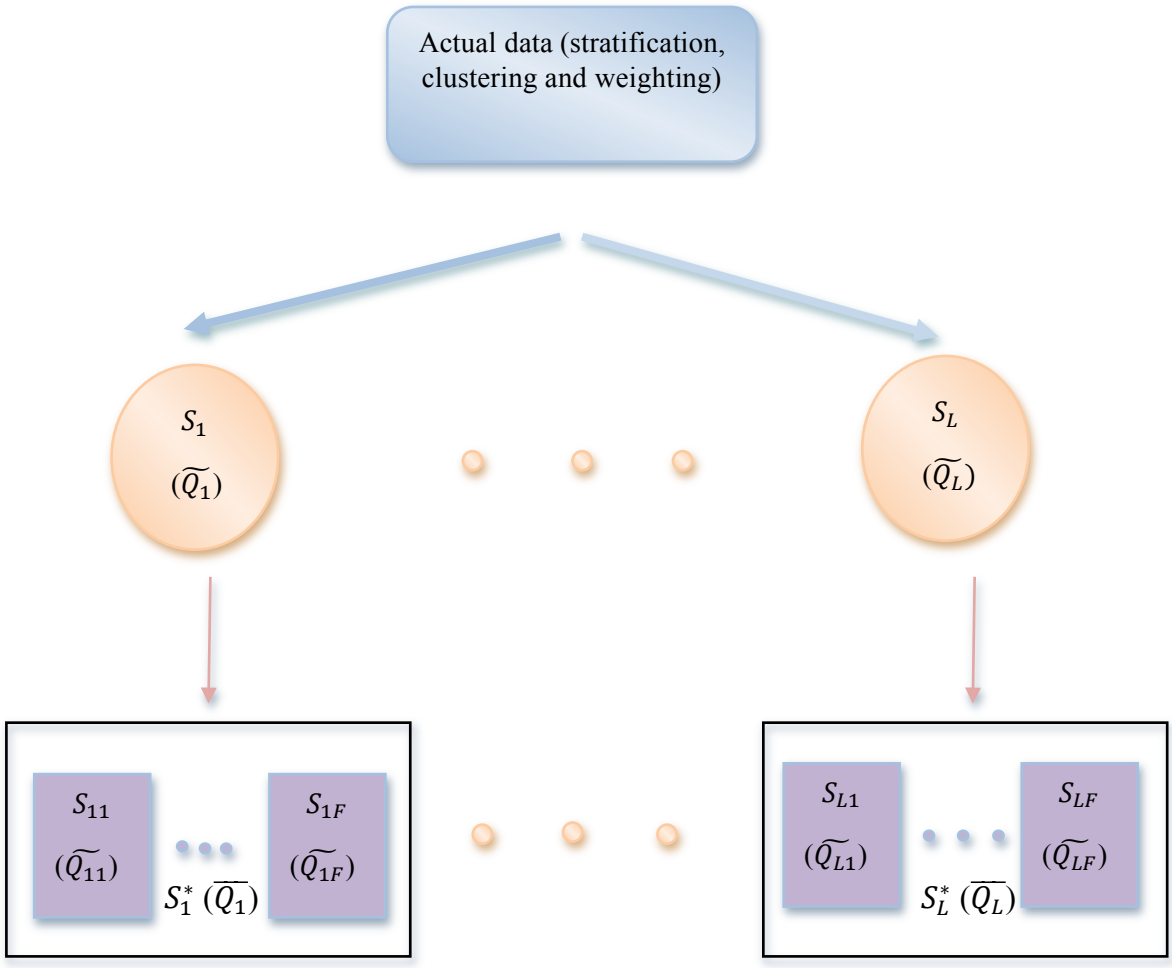
Once we have *L* BB samples with replicate weights, the second step imputes the unobserved units using the FPBB Polya urn scheme (Theorem 1). In practice, the probability of selecting the k^{th} unit, y_k^*, depends on the selection of the first k-1 units, $y_1^*, \dots, y_{k-1}^*$. In another words, to determine the probability of selecting a new unit, we have to count the number of times that each unit in the sample has been selected among

the previous selections. To make it computationally more efficient, we could draw a moderate size of population for multiple times and then pool them to simulate the posterior predictive distribution.

Assume F FPBB samples of size $k * n$ are produced for each BB sample, denoted by $S_{l1}, \dots, S_{lF}, l = 1, \dots, L$, where k is an integer. We pool the F FPBB samples to produce one synthetic population, S_l^*. The size of S_l^* then is $F * k * n$.

Figure 3.2 provides a flowchart that summarizes this procedure. $S_l^*,\ l = 1, \dots, L,$ are the synthetic populations, which can be analyzed using the standard statistical methods for simple random samples. We denote the estimate of the statistic of interest Q from S_{lf} by $\widetilde{Q_{lf}}, l = 1, \dots, L, f = 1, \dots, F$ and $\overline{Q_l} = \frac{\sum_{f=1}^{F} \widetilde{Q_{lf}}}{F}, l = 1, \dots, L$ is the estimate obtained from the synthetic population, S_l^*. Inference can be directly made from the synthetic data combining rule (Raghunahtan *et al* 2003), which is essentially the same as the combining rule for the bootstrap samples when the number of synthetic populations is big.

Figure 3.2 Nonparametric method to impute the unobserved population



3.4 Randomization Validity

In this section, we evaluate the performance of the point estimate and variance estimate from the nonparametric method from the randomization perspective. Assume we generate L synthetic populations, $S_l^*,\ l = 1, \dots, L$. The estimate of the population quantity of Q obtained from from S_l^* is denoted by $\overline{Q_l}$. Raghunathan *et al.* (2003) suggest to estimate Q by $\overline{Q_L} = \frac{\sum_{l=1}^{L} \overline{Q_l}}{L}$ with variance estimate $Var(\overline{Q_L}) = \frac{1}{L}\sum_{l=1}^{L}(\overline{Q_l} - \overline{Q_L})^2$.

We first prove $\overline{Q_L}$ is unbiased for Q. Recall that S_l^* is generated by pooling the F synthetic populations that impute the unobserved units of $S_l, l = 1, \dots, L$. From Theorem 1, $E\left(\widetilde{Q_{lf}} \left| S_l\right.\right) = \widetilde{Q_l}, l = 1, \dots L, f = 1, \dots, F$. From the well established bootstrap theories, $E\left(\widetilde{Q_l}\middle| S\right) = q$, where q is the estimate obtained from the actual data. If q is estimated after adjusting for the complex sampling design features, $E(q|\mathcal{P}) = Q$, where q denotes the true population. Thus,

$E(\overline{Q_L}|\mathcal{P}) = E\left(E\left(E\left(\frac{\sum_{l=1}^{L}\overline{Q_l}}{L}\left| S_l\right.\right)\right| S\right)|\mathcal{P}) = E\left(E\left(E\left(\frac{\sum_{l=1}^{L}\frac{\sum_{f=1}^{F}\widetilde{Q_{lf}}}{F}}{L}\left| S_l\right.\right)\right| S\right)|\mathcal{P}) =$

$E(E(\frac{\sum_{l=1}^{L}\widetilde{Q_l}}{L}|S)|\mathcal{P}) = E(q|\mathcal{P}) = Q.$

The variance of $\overline{Q_L}$ is estimated by the between variance $\frac{1}{L}\sum_{l=1}^{L}(\overline{Q_l} - \overline{Q_L})^2$. $var(\overline{Q_{boot}}) = \frac{1}{L}\sum_{l=1}^{L}(\widetilde{Q_l} - \overline{Q_{boot}})^2$, where $\widetilde{Q_l}$ is calculated using the replicate weights and $\overline{Q_{boot}} = \frac{1}{L}\sum_{l=1}^{L}\widetilde{Q_l}$. From the bootstrap theory, $var(\overline{Q_{boot}}) = var(q|\mathcal{P})$ is the variance estimate of q after the complex sampling design features are adjusted for. In the second step when we generate L synthetic populations, $S_l^*, l = 1, \dots, L$, which can be analyzed as simple random samples. If F and k are infinite, the estimates from S_l^*, $\overline{Q_l} = \frac{\sum_{f=1}^{F}\widetilde{Q_{lf}}}{F}$, is unbiased for $\widetilde{Q_l}$. Then the variance estimates of Q obtained from the synthetic populations, $var(\overline{Q_L}) =$ $\frac{1}{L}\sum_{l=1}^{L}(\overline{Q_l} - \overline{Q_L})^2 = \frac{1}{L}\sum_{l=1}^{L}(\widetilde{Q_l} - \overline{Q_{boot}})^2 = var(\overline{Q_{boot}}) = var(q|\mathcal{P})$, which are the variance estimates after complex sampling design features are taken into account.

In practice, it is not realistic to set F and k to be infinite, which may result in a random error for the estimate of Q from the synthetic populations. Assume we have the

following random measurement error model, $\overline{Q_l}=\widetilde{Q_l}+e_l, l=1,\dots,L$ and $e\sim N(0,\sigma_e^2)$, i.e., $\frac{1}{\mathrm{L}}\sum_{l=1}^{L}e_l=0, \frac{1}{\mathrm{L}}\sum_{l=1}^{L}e_l^2=\sigma_e^2$. This brings in extra variability into the variance estimate. Under the assumed measurement error model,

$$\frac{1}{L}\sum_{l=1}^{L}(\overline{Q_l}-\sum_{l=1}^{L}\overline{Q_l}/L)^2=\frac{1}{L}\sum_{l=1}^{L}(\widetilde{Q_l}+e_l-\overline{Q_{boot}})^2=\frac{1}{L}\sum_{l=1}^{L}(\widetilde{Q_l}-\overline{Q_{boot}})^2+$$

$$\frac{1}{L}\sum_{l=1}^{L}{e_l}^2+\frac{1}{L}2\sum_{l=1}^{L}\widetilde{Q_l}e_l.$$

Considering $\widetilde{Q_l}, l=1,\dots,L$ are the estimates obtained from L independent draws and $e_l\sim N(0,\sigma_e^2)$, the last term $\frac{1}{L}2\sum_{l=1}^{L}\widetilde{Q_l}e_l$ should be trivial. Thus, the variance estimate from the synthetic populations is $var(\overline{Q_L})=\frac{1}{L}\sum_{l=1}^{L}(\widetilde{Q_l}-\overline{Q_{boot}})^2+\sigma_e^2$. σ_e^2 can be made arbitrarily small by increasing the synthetic population size or increasing the number of FPBB draws F. From our simulation studies, we suggest the minimum F and k are $F=5$ and $k=5$.

3.5 Simulation Study

In this section, we conduct a simulation study to evaluate the nonparametric method that generates synthetic populations while adjusting for the complex sampling design features. We use a simulated population in the study so that we can evaluate the repeated sampling properties of the nonparametric method.

We create a population with strata and clusters within each stratum from the following bivariate normal distribution:

$$\begin{pmatrix}X_{1ijk}\\X_{2ijk}\end{pmatrix}\sim N\left(\begin{pmatrix}500+4.5*i+u_{ij}\\500+4.5*i+u_{ij}\end{pmatrix},\begin{pmatrix}100 & 50\\50 & 100\end{pmatrix}\right),$$

where $i=1{:}\,150$ denotes the stratum effect,